ANALYSIS OF DATA IN TIME

PROCEEDINGS OF THE 1989 INTERNATIONAL SYMPOSIUM

Edited by
A.C. Singh and P. Whitridge

Statistics Canada
Statistique Canada

Canada

ANALYSIS OF DATA IN TIME

METHODOLOGY SYMPOSIUM SERIES

1984	Analysis of Survey Data
1985	Small Area Statistics
1986	Missing Data in Surveys
1987	Statistical Uses of Administrative Data
1988	The Impact of High Technology on Survey Taking
1989	Analysis of Data in Time
1990	Measurement and Improvement of Data Quality (forthcoming)

ANALYSIS OF DATA IN TIME

Proceedings of the 1989 International Symposium

Edited by

A.C. Singh and P. Whitridge
Methodology Branch
Statistics Canada
Ottawa, Ontario, Canada

Statistics Canada Statistique Canada

Canada

Price: $20.00 Canadian

Payable to

"The Receiver General for Canada - Symposium '89 Proceedings"

Send your request to:

Statistics Canada Symposium '89 Proceedings
Social Survey Methods Division
R.H. Coats Building
Tunney's Pasture
Ottawa, Ontario
Canada
K1A 0T6

Preface

In recent years, there has been a growing demand within government and private sectors for statistical tools suitable for analysing data collected periodically over time from sample surveys, censuses and administrative sources. In view of this demand, an international symposium on Analysis of Data in Time was organized to bring together researchers and practitioners in various substantive fields from universities, government and other statistical agencies. It was sponsored by Statistics Canada and the Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa.

The symposium was held October 23-25, 1989 in the Simon Goldberg Conference Centre at Statistics Canada, Ottawa, attended by about 325 registered participants. Several papers from well known statisticians around the world were presented. The key note address was given by Prof. Wayne Fuller of Iowa State University. The special invited lecture by Prof. David Brillinger of University of California at Berkeley could not be presented at the symposium due to the difficult circumstances caused by the earthquake in California. It is nevertheless included in the proceedings for the benefit of readers.

The present volume contains 27 papers with varying levels of theoretical and applied content. It is believed that the wide range of topics covered in the symposium would be very useful to both researchers and practitioners engaged in various fields of statistics. The papers have been organized into the following eight parts:

Part 1:	Sampling on Repeated Occasions
Part 2:	Time Series Analysis in the Presence of Survey Error
Part 3:	Analysis of Time Series of Counts
Part 4:	Developments in the Analysis of Time Series Data
Part 5:	Epidemiology
Part 6:	Demography
Part 7:	Econometrics
Part 8:	Education

The Proceedings also includes the opening remarks given by G. Brackstone and the closing remarks by D. Binder. The French translations of the papers were reviewed by a number of methodologists. Our sincere appreciation goes to J. Armstrong, S. Beaulieu, J.-M. Berthelot, J.-R.Boudreau, R. Boyer, M. Brodeur, M. Bureau, P. Daoust, P. David, J. Denis, J. Dufour, J. Dumais, S. Giroux, M. Joncas, M. Lachance, D. Lalande, E. Langlet, Y. Leblond, J. Lynch, S. Perron, C. Morin, C. Poirier, G. Sampson, P. St-Martin, A. Théberge, M. Thibeault, and J. Tourigny. It is also our great pleasure to thank Judy Clarke, Carole Jean-Marie, Christine Larabie, Carmen Lacroix and Pat Pariseau for their efficient manuscript processing and especially Judy for coordinating the production work.

The organization of the symposium was supported by many persons at Statistics Canada, especially J. Mayda and J. Morabito. We would also like to thank D. Binder, G. Brackstone, D. Drew, J. Kovar, J.N.K. Rao, and M.P. Singh for their encouragement and consultation. Finally, our appreciation must be offered to the speakers for making the symposium a great success.

A.C. Singh
P. Whitridge
Organizing and Editorial Committee
Symposium '89

Ottawa, Ontario Canada
October 1990

TABLE OF CONTENTS

OPENING REMARKS

PART 1. SAMPLING ON REPEATED OCCASIONS

PART 2. TIME SERIES ANALYSIS IN THE PRESENCE OF SURVEY ERROR

PART 3. ANALYSIS OF TIME SERIES OF COUNTS

PART 4. DEVELOPMENTS IN THE ANALYSIS OF TIME SERIES DATA

PART 5. EPIDEMIOLOGY

PART 6. DEMOGRAPHY

PART 7. ECONOMETRICS

PART 8. EDUCATION

CLOSING REMARKS

CONTRIBUTORS

J.R. BALDWIN, Business and Labour Market Analysis Group, Statistics Canada, Ottawa, Ontario, Canada, and Department of Economics, Queen's University, Kingston, Ontario, Canada.

W.R. BELL, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

D.R. BELLHOUSE, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

S. BEN-TUVIA, Central Bureau of Statistics, Jerusalem, Israel.

D.A. BINDER, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

G.J. BRACKSTONE, Informatics and Methodology Field, Statistics Canada, Ottawa, Ontario, Canada.

D.R. BRILLINGER, Department of Statistics, University of California, Berkeley, California, U.S.A.

L. BURCK, Central Bureau of Statistics, Jerusalem, Israel.

R.T. BURNETT, Environmental Health Centre, Health and Welfare Canada, Ottawa, Ontario, Canada.

P.A. CHOLETTE, Time Series Research and Analysis Division, Statistics Canada, Ottawa, Ontario, Canada.

G.H. CHOUDHRY, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

W. CLARK, Education, Culture and Tourism Division, Statistics Canada, Ottawa, Ontario, Canada.

J.P. DICK, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

Q.P. DUONG, Bureau of Management Consulting, Ottawa, Ontario, Canada.

D.F. FINDLEY, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

C. FORTIER, Demography Division, Statistics Canada, Ottawa, Ontario, Canada.

W.A. FULLER, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.

K. FYFE, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

J.F. GENTLEMAN, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada.

P.K. GORECKI, Economic Council of Canada, Ottawa, Ontario, Canada.

R.M. HARTER, A.C. Nielsen, Northbrook, Illinois, U.S.A.

V.K. JANDHYALA, Department of Mathematics, Washington State University, Pullman, Washington, U.S.A.

H. JOHANSEN, Health Promotion Directorate, Health and Welfare Canada, Ottawa, Ontario, Canada.

J.D. KALBFLEISCH, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

L. KISH, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, U.S.A.

D. KREWSKI, Environmental Health Centre, Health and Welfare Canada, Ottawa, Ontario, Canada.

N. LANIEL, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

J.F. LAWLESS, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

P. LIN, Industry Measures and Analysis Division, Statistics Canada, Ottawa, Ontario, Canada.

L. LIU, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

I.B. MacNEILL, Department of Statistical and Actuarial Sciences and Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada.

R.H. McGUCKIN, Center for Economic Studies, U.S. Bureau of the Census, Washington, D.C., U.S.A.

B.C. MONSELL, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

C. NAIR, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario, Canada.

M. NARGUNDKAR, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

T. PETERSEN, Industry Measures and Analysis Division, Statistics Canada, Ottawa, Ontario, Canada.

D. PFEFFERMANN, Hebrew University, Jerusalem, Israel.

R. PRESSAT, Département de la Conjoncture, Institut national d'études démographiques, Paris, France.

M.G. PUGH, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.

J.N.K. RAO, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

G.R. ROBERTS, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

D. ROBERTSON, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada.

G. ROWE, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada.

L. SAGER, Industry Measures and Analysis Division, Statistics Canada, Ottawa, Ontario, Canada.

J. SHEDDEN, Environmental Health Centre, Health and Welfare Canada, Ottawa, Ontario, Canada.

A.C. SINGH, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada.

J. STRACHAN, School of Health Information Science, University of Victoria, Victoria, British Columbia, Canada.

T. WANNELL, Business and Labour Market Analysis Group, Statistics Canada, Ottawa, Ontario, Canada.

M.C. WOLFSON, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada.

K.M. WOLTER, A.C. Nielsen, Northbrook, Illinois, U.S.A.

OPENINGS REMARKS

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

INTRODUCTION

G.J. Brackstone[1]

On behalf of Statistics Canada may I welcome you to Symposium 89. This symposium is sponsored jointly by Statistics Canada and the Laboratory for Research in Statistics and Probability of Carleton University and the University of Ottawa. It is very encouraging to see such a large crowd here this morning. It shows that either we have picked a very pertinent topic and designed an attractive program, or our Organizing Committee has undertaken a very successful marketing effort, or both.

The theme of this Symposium is the Analysis of Data in Time. This title, of course, has a certain ambiguity about it, at least in English. Those of you who have come here to learn how you can speed up your analysis, or ensure that it meets deadlines, may be in for disappointment because that is not the sense in which we have interpreted this year's theme. It is with the collection, processing, and especially the analysis of data in the time dimension that we shall be concerned.

In keeping with this theme let me say that this Symposium is the sixth realization in the time series of methodology symposia at Statistics Canada. Previous symposium topics from 1984 to 1988 have been: Analysis of Survey Data (1984) - where we focussed on cross-sectional analysis of data from complex surveys; Small Area Statistics (1985) - which resulted in a published book; Missing Data in Surveys (1986) - a smaller symposium but with some top-notch speakers on this problem area for statistical agencies; Statistical Uses of Administrative Data (1987) - where we had a really international set of speakers on both statistical and privacy aspects of this topic; and then last year, The Impact of High Technology on Survey Taking - where we explored the synergy between survey methodology and informatics.

For enthusiasts, I'll leave the question of whether this symposia time series is a random or non-random realization. For those who specialize in prediction, I leave the challenge to predict the topic for next year's symposium before it is announced later in the week.

At Statistics Canada we believe that these symposia have many benefits - otherwise we wouldn't persist with them. They provide the opportunity for theoreticians and practitioners to come together to discuss a topic of real and practical importance to statistical agencies. They serve, we hope, to generate interest among statisticians outside statistical agencies in applications of importance to statistical agencies. They provide a focus and deadline for both our staff and statisticians outside to complete relevant research work, and to exhibit it for peer review. And for our own staff, they provide the opportunity to listen to some of the world's foremost statisticians without having to get travel approval.

The choice of this year's topic, the Analysis of Data in Time, is to my mind both timely and appropriate. It provides a forum for the exchange of ideas between theorists and practitioners and between statisticians from universities and those from governments and other agencies. Despite important developments in time series theory and methods, and notwithstanding the availability of data generated by repeated experiments, regular surveys, censuses and administrative files, there are time series methods with well known and worthwhile features which, far from being in routine use, are almost never used in government agency programs.

There are perhaps three main factors which have brought about this situation:

First, these methods often involve rather complex calculations and data handling, and a fairly heavy load of computations;

Second, practitioners, especially those carrying out the investigations, may be unfamiliar with the current theory;

And third, there are undeniable weaknesses and deficiencies in the theory -- it does not cover all the situations faced by practitioners.

The first of these causes: computational complexity, while not to be dismissed outright is not so important an issue today and is likely to be even less important in the future. But the other two causes: lack of familiarity with theory on the part of practitioners, and shortcomings in theory, will persist unless we do something about them. And that is why we are having this Symposium. It is one of the ways in which we try to bridge the gap between theory and practice, between theoreticians and practitioners.

[1] G.J. Brackstone, Informatics and Methodology Field, Statistics Canada, Ottawa, Ontario K1A OT6

To flourish, the community of theoreticians needs real and important problems on which to work. Practitioners can offer this. Much of a practitioner's work is concerned with the tailoring and implementation of theory for specific applications, in the course of which limitations of existing theory may be discovered, thus providing further challenges for the theorist.

Now I want to say a few words about why this topic is important to Statistics Canada, as well as to other statistical agencies. Almost all the data we publish are time series. There are not many statistics for which one could say that the only interest is in its value today. People want to know how things are changing, and that means time series - whether we call them that or not. So what are the trends that make this Symposium's theme particularly important?

Like everyone else we are suffering resource constraints. We therefore want to extract the maximum information out of existing data without additional costly data collection. Bringing the time dimension into the analysis can help in this regard.

As our primary source of data, the design of surveys has to be optimized. More attention to the time dimension in both the design and estimation stages of surveys whose data will be used to monitor change may yield significant benefits.

Another prime concern is user understanding and interpretation of data we publish. Some of the least understood aspects of our data are time-related. I refer to seasonal adjustment and to revision practices that incorporate later data into series published earlier in preliminary form. We believe there is progress to make, if not in simplifying such procedures, at least in explaining them, and in ensuring that they result in consistent data sets.

Finally, there is the growing interest in longitudinal data at the micro-level - that is, information about how individuals (persons, businesses, farms, etc.) are changing, rather than only how the aggregate measures are changing. Here again time series methods can help.

These are some of the issues we are facing today that make the theme of this Symposium an important one for us.

The program looks to me like a very interesting one with a good mixture of theory and practice in a variety of fields including demography, econometrics, education and epidemiology. I hope that each of you will benefit from this Symposium, and that at least some of you will be inspired to pursue further the development or application of theory in this area. I hope also that some interest may be generated in more collaborative work between university and government statisticians.

Thank you all for supporting this Symposium and I wish you an interesting and productive three days.

PART 1

SAMPLING ON REPEATED OCCASIONS

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

ANALYSIS OF REPEATED SURVEYS

Wayne A. Fuller[1]

ABSTRACT

KEY WORDS: Survey sampling, least squares, measurement error, gross change.

Repeated surveys in which a portion of the units are observed at more than one time point and some units are not observed at some time points are of primary interest. Least squares estimation for such surveys is reviewed. Included in the discussion are estimation procedures, modified so that existing estimates are not revised when new data become available. Also considered are techniques for the estimation of longitudinal parameters, such as gross change tables. Estimation for a repeated survey of land use conducted by the U. S. Soil Conservation Service is described. The effects of measurement error on gross change estimates is illustrated and it is shown that a survey design that estimates the parameters of measurement error process can be very efficient.

1. INTRODUCTION

There is considerable interest in the analysis of surveys that are repeated in time. Evidence of this interest are the recently published proceedings of a conference on panel surveys edited by Kasprzyk, Duncan, Kalton and Singh (1989), sessions at the last two meetings of the International Statistical Institute, and this conference. Smith and Holt (1989) at the 1989 ISI session in Paris call this a "resurgence of interest in the design and analysis of longitudinal studies." They note that researchers in areas such as sociology and health have long conducted panel surveys and cohort studies. They cite, as an example, Lazarsfeld and Fiske (1938). An example in a health related area is Garcia, Battese, and Brewer (1975).

Official agencies conduct many surveys, such as labor force surveys, on a regular basis. The output of such surveys is usually a sequence of reports, such as those on current employment and unemployment. Typically, very few statistics on the behavior of individual units over time have been reported from repeated official surveys. An example of a survey designed to produce longitudinal estimates is the U.S. Survey of Income and Program Participation. See Kasprzyk and McMillen (1987). While information on private surveys is less complete than that on government surveys, it seems that the most common use of repeated private surveys is also to produce a sequence of reports for points in time. However, the demand for longitudinal analysis has increased for both public and private data providers.

The complex issues associated with repeated surveys are brought into focus when one attempts to develop a taxonomy for such studies. Duncan and Kalton (1987) list some seven objectives of surveys repeated over time. These are:

A. To provide estimates of population parameters at distinct time points.
B. To provide estimates of population parameters summed across time.
C. To measure net change at the aggregate level.
D. To measure components of change including
   i) gross change
   ii) change for an individual
   iii) variability for an individual
E. To aggregate individual data over time.
F. To measure the frequency, timing and duration of events.
G. To accumulate information on rare populations.

While not mentioned explicitly, several of these objectives implicitly include the estimation of the parameters of subject matter models.

Duncan and Kalton also define four kinds of surveys. Their definitions were: (1) repeated survey, in which no attempt is made to guarantee that particular elements appear in more than one sample; (2) the pure panel survey, in which the same elements are observed at every point in time; (3) the rotating panel survey, in which there is a fixed pattern under which

[1] Department of Statistics, Iowa State University, Ames, Iowa, 50011.

elements are observed for a fixed number of times and then rotated out of the sample; and (4) the split panel survey, in which a pure panel survey is combined with a repeated survey or a rotating panel survey. Duncan and Kalton present a table in which they outline how the different kinds of surveys are appropriate for the different kinds of objectives.

An institution conducting a repeated survey faces all of the usual survey problems, but the problems are magnified. Nonresponse is always a concern, but it is more difficult to maintain cooperation over a period of time. Response error is always present, but repeated surveys encounter problems of "conditioning" associated with repeated interviews. Also, response errors introduce inconsistencies into data collected over time. The quality repetition of a survey requires maintaining consistent field, processing, and estimation procedures over time. Also data management problems increase for repeated surveys. Finally, the changing composition of units, such as families, over time complicates estimation and analysis.

We shall examine only a few issues associated with repeated surveys. Our discussion is motivated by a large scale survey conducted by the U.S. Soil Conservation Service with the cooperation of Iowa State University. In Section 2 we review some of the estimation techniques applicable for repeated surveys. This discussion is continued in Section 3 with more emphasis on estimation of longitudinal parameters in panel surveys. In Section 4 we briefly describe the estimation procedures used in the U.S. Soil Conservation Service study. Section 5 contains a short description of the effects of measurement error on gross change estimates.

2. ESTIMATION

In this section we outline generalized least squares estimation for surveys with only a subset of elements observed at successive times. Generalized least squares was the procedure first considered by authors studying estimation for surveys repeated in time. Beginning with Jessen (1942), who was influenced by Cochran (1942), authors considered the construction of minimum variance weights for a set of unbiased estimators available at each point in time of the survey.

Jessen (1942) investigated the special case of sampling on two occasions with unequal numbers of observations, and studied the optimal allocation of units to overlapping and nonoverlapping sample groups. Patterson (1950) considered sampling on T occasions under several schemes of partial replacement of units. The simplest such sampling plan required the replacement of a fixed proportion of sampling units on each successive sampling occasion. Also, Patterson (1950) assumed that for a given i, the differences $x(ti) - x(t)$, $t=1, 2, \ldots,$ followed a first-order autoregressive process, where $x(ti)$ was the value of the i-th population unit at time t, and $x(t)$ was the corresponding finite population mean. Under the resulting error model, he developed optimal estimators of the fixed $x(t)$ values and of the differences $x(t) - x(t-1)$. He also considered the optimal estimation of $x(t)$ under generalizations of the partial replacement plan, optimal sample size selection, and estimation with nonautoregressive errors.

Least squares procedures were considered further by Eckler (1955), Gurney and Daly (1965) and Jones (1980). Composite estimation was a name given to certain types of estimators. See Rao and Graham (1964), Graham (1973), and Wolter (1979). Battese, Hasabelnaby and Fuller (1989) describe the application of the least squares procedure to the farm survey conducted by the U.S. Department of Agriculture.

It seems fair to say that the parameters under consideration by these authors were means or totals at specific time points. That is, longitudinal parameters, such as the fraction of individuals in a particular class at both time 1 and time 2, were not explicitly considered by these authors. However, as we shall see, the least squares method extends to such parameters.

Linear least squares has the desirable feature that estimators for a number of characteristics are internally consistent. That is, the least squares estimator of $\bar{Y}$ plus the least squares estimator of $\bar{Z}$ is the least squares estimator of $\bar{Y} + \bar{Z}$. However, if different vectors of observations are used to construct different estimates, the internal consistency is destroyed.

In many applied surveys it is not possible to compute the optimum least squares estimators for all points in time. First, all available information can not be used in the estimation. That is, it is not possible to incorporate all data from the surveys of preceding times into a least squares analysis for the current time. Often the number of variables exceeds the number of observations. Second, the releasing organization may be restricted in the number of times they revise previous estimates. This second point has been discussed by Smith and Holt (1989).

To illustrate these estimation problems, we have constructed a small example. The example two-way table for classification at two points in time, as observed in a very large sample, is given in Table 1. We have given names to this table, letting the first category be employed and letting the second category be unemployed. We shall assume that the population is constant over time. If there are births and deaths, then the table would need to be increased to a 3 × 3 table. Let us assume that we are interested in estimating the change in level from one period to the next. Let us also assume that we are interested in the gross change table which involves estimating the interior cells of the table. In the 2 × 2 table it is only necessary to estimate the (1, 1) cell and the marginal proportions to define all cells of the table.

Table 1. Hypothetical proportions for two points in time

	TIME 2		
TIME 1	Employed	Unemployed	Total
Employed	0.91	0.02	0.93
Unemployed	0.03	0.04	0.07
Total	0.94	0.06	1.00

We assume a two period study in which an equal number of elements are observed at each of the two times. We assume that one-half of the elements observed at the first time are also observed at the second time. That is, of the elements observed at the second time, one-half were observed at the first time and one-half are new to the sample. We take as our vector of observations the proportion of elements in category 1 in the one-half of the sample that is not observed the second time [denoted by P(E.1)], the proportion of elements in category 1 at time 1 in the remaining half of the sample [denoted by P(E.2)], the elements that are in category 1 at both time 1 and time 2 for the portion of the sample that is observed at both time periods [denoted by P(EE)], the proportion of the elements in category 1 at time 2 for the elements that are observed at both times [denoted by P(.E2)], and the proportion of elements in category 1 at time 2 for the portion of the sample that is observed only at time 2 [denoted by P(.E3)]. We shall place arguments in parenthesis when the expressions appear in the text and place the arguments as subscripts in the displays.

We assume simple random sampling. Then, because the statistics are sample proportions, it is easy to write down the covariance matrix of the vector of five estimates. A multiple of that covariance matrix is given in Table 2. To obtain the covariance matrix for a sample of size n at each time period, divide every entry in the table by n and multiply by two. In Table 3 we give the variance of alternative estimation procedures. In the first column is the variance of the procedure that uses as the estimator of the first period proportion only the elements appearing in the first period sample. To estimate the fraction appearing in category 1 (employed) both at time 1 and time 2, the simple procedure uses only the overlap

Table 2. Covariance matrix of the vector of sample proportions, two time points and fifty percent overlap in sample. (For a sample of size n multiply entries by 2 and divide by n .)

$P_{E.1}$	$P_{E.2}$	P_{EE}	$P_{.E2}$	$P_{.E3}$
0.0651	0	0	0	0
0	0.0651	0.0637	0.0358	0
0	0.0637	0.0819	0.0546	0
0	0.0358	0.0546	0.0564	0
0	0	0	0	0.0564

elements, and to estimate the number in the first category at time t , it uses only the sample observed at time 2. Thus, if we have a sample of 200 elements at each time period, the first period sample of 200 elements is used to estimate the first probability. The 100 elements observed at both time 1 and time 2 are used to estimate the elements staying in category 1, and the 200 elements observed at time 2 are used to estimate the time 2 proportion.

Table 3. Variance of alternative estimation procedures (For a sample of size n at each period, multiply entries by 2 and divide by n .)

	Procedure		
Parameter	Simple	Restricted GLS	Full GLS
$P_{E.}$	0.0326	0.0326	0.0294
P_{EE}	0.0819	0.0397	0.0374
$P_{.E}$	0.0278	0.0258	0.0255
$P_{EE}/P_{.E}$	0.0290	0.0229	0.0220
$P_{.E} - P_{E.}$	0.0429	0.0367	0.0353

The last column is the variance of the best linear unbiased estimators constructed using generalized least squares. The estimators are constructed from the vector of five basic statistics and the covariance matrix of that vector. This estimator is of the form

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \ , \qquad (1)$$

where V is given in Table 2, $\beta = (P_{E.}, P_{.E}, P_{EE})$,

$$X' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} ,$$

and Y is the five-dimensional vector of direct estimates,

$$Y' = (\tilde{P}_{E.1}, \tilde{P}_{E.2}, \tilde{P}_{EE}, \tilde{P}_{.E2}, \tilde{P}_{.E3}) \ .$$

The second column of Table 3 gives the variance of the restricted least squares estimators, where the restriction is that the estimator for the first period must be the estimator obtained from the initial sample. This would be the appropriate procedure if the agency never made a revision in the once published estimates. For example, the Bureau of Labor Statistics in the United States does not revise the unemployment statistics. Once released, they are the official estimates. Of course, the United States unemployment statistics are based on a more complicated sample and are based on a survey that is conducted over a longer period of time.

To describe the restricted generalized least squares estimator of Table 3, let the model be

$$Y = X\beta + e \ ,$$

where X is a fixed $n \times k$ matrix and

$$E(ee') = V \ .$$

The generalized least squares estimator of β , with some elements of β restricted to be certain liner combinations of Y can be constructed as follows. Consider the Lagrangian

$$(Y - X\beta)' \ V^{-1}(Y - X\beta) - 2 \sum_{i=1}^{b} \lambda_i(\Gamma_i\beta - g_i)$$

where $\Gamma(i)$ is a fixed row vector and b is the number of restrictions. The solution to this minimization problem is defined by

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X'V^{-1}Y \\ g \end{pmatrix},$$

where $\lambda' = (\lambda_1, \lambda_2, \ldots, \lambda_b)$, $\Gamma' = (\Gamma_1', \Gamma_2', \ldots, \Gamma_b')$ and $g' = (g_1, g_2, \ldots, g_b)$.

If we replace g by the linear combination GY , the equation becomes

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} Y \ .$$

This equation defines the restricted estimator of β as a linear function of Y . Hence the variance of the estimator of β is the upper $k \times k$ portion of

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} V \left[\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} \right]'$$

This is not the only way to compute the restricted generalized least squares estimator. An alternative estimator of level and change that leaves the previous estimator unchanged is the composite estimator. See, for example, Wolter (1979).

Several points are illustrated by this small example. First, with a correlation of 0.591 between employment at the two time periods, the improvement in the current estimate of unemployment from using generalized least squares is modest, about 10%. On the other hand, there is a very large improvement in the variance of the estimate of P(EE) from using generalized least squares. The variance of the generalized least squares estimator is about 45% of the variance of the simple estimator. The second important point is that the use of restricted generalized least squares to estimate P(EE) and P(.E) produces estimates that are nearly as efficient as full generalized least squares. There is about a one percent loss for the estimate of P(.E) and about a six percent loss for the estimate of P(EE) .

3. LONGITUDINAL ESTIMATORS

Recall that our definition of a pure panel survey is one in which the same elements are observed at every time point of data collection. The pure panel survey is possible for observations of certain physical units, such as plots of land. In the case of surveys of human populations, the pure panel must be classed as a figment of the statistician's imagination. In the real world, a fraction of the respondents from the first time are always unavailable at the second time. Good reviews of procedures for missing data are given by Lepkowski (1989) and Little and Su (1989). Also see Little and Rubin (1987), Kalton (1983), and Madow et al. (1983).

We have described the rotating panel survey in which the design calls for some elements to leave the study and some elements to enter the study at every time point at which the study is conducted. In this type of survey we might say that we have planned nonresponse for those elements that are rotated out of the sample. Thus, estimation in the presence of nonresponse and estimation for rotating panel surveys are related problems.

Given that one does not obtain data from every respondent at every point in time of a repeated survey, one is faced with a choice among methods of handling planned and unplanned nonresponse. There are two simple, and common, procedures. If the interest is in following individuals over time, then very often the investigator retains in the study only those individuals that responded every time. A weighting procedure may be used to adjust the data using characteristics of the initial respondents and (or) external auxiliary data. This procedure is often used in special one-time studies of a specific population. In such situations the report on the study is released only after the entire study is completed.

The second common type of estimation procedure is to construct estimates for each time period using the data that are available for that time period. This procedure is often used if the survey is repeated regularly, the results are released after each survey, no revisions are made in the releases, and no longitudinal estimates are produced. One-period-at-a-time estimation has the advantage of being very easy to compute at time t because no information from the previous period is used in calculating the current estimators. It generally gives good estimates (not optimal) of the current value, but rather poor estimates of change.

In fact, one might use both of these procedures in a single survey. The Survey of Income and Program Participation (SIPP) conductd by the U.S. Bureau of the Census is a panel survey with a rotating time- of-interview with a four month recall period. The Census Bureau provides a set of weights at each time of the survey that can be used to construct estimates for that point in time using all individuals that respond at that time point. They also provide (a) the sample of individuals that responded all eight times for the period 1984-85 with weights for these individuals, (b) the sample of individuals that resonded all four times in 1984 with an appropriate weight and (c) the sample of individuals that responded all four times in 1985 and an appropriate weight.

We outline an estimation procedure for a panel survey with nonresponse where the analysis is conducted at the end of the survey. It is assumed that a reasonable fraction of the units respond at all time points of the survey and that longitudinal analysis is of interest. The computational procedure consists of constructing weights for the units with complete response records. Information from respondents with incomplete records constitutes a form of auxiliary information.

The first step in the analysis is to pick a few variables that are very important to the study. The number of variables that can be used will depend upon the sample size. The covariance structure of the vector of estimates composed of the simple estimates for each of these variables for each type of response pattern for each point in time where the estimate is appropriate, is computed. The covariance structure is a function of the response-nonresponse pattern. There are different definitions of simple estimators. For simple random sampling, simple estimators are simple means. For stratified samples, one might define the original vector to include estimates for each stratum. Alternatively, the simple estimator for a stratified sample might weight the responses in each stratum for nonresponse. The vector $\mathbf{Y}$ used in (1) is an example of a vector of simple estimates.

Given the vector of simple estimators and the estimated covariance matrix of the vector, improved estimators for each of the time periods is constructed by generalized least squares. For example, if we had a panel study with three time points, there are seven response patterns. These are XXX, OXX, XOX, XXO, XOO, OXO, OOX, where X denotes response and O denotes nonresponse. If we choose two variables of interest, the vector of simple estimates will contain $12 \times 2 = 24$ estimates because there are 12 group-response times associated with the seven response patterns. In this example, generalized least squares would be used to produce six estimates, the estimates for the two variables for each of the three time periods.

The generalized least squares estimator for the selected charcteristics become control variables for a next stage of estimation. Using regression weighting methods, weights are constructed for the individuals that responded all times. The weights are constructed so that the generalized least squares estimates for each time period are reproduced by the weighted sample of 100% respondents. That is, the time estimates for the chosen variables are used as controls.

The efficiency of this procedure depends upon the correlation between the chosen control variables and the analysis variable. If a control variable is also the analysis variable, the procedure will be very efficient. It is less than fully efficient only because a limited amount of information is used in the generalized least squares procedure.

The strong advantage of this procedure is that it produces a single tabulation data set that can be used to construct internally consistent estimates for all reporting times and for all gross change tables.

The variance of the procedure can be computed by analogy to the procedures used for double sampling. Let Y be the characteristic of interest. For simplicity, assume a simple random sample at each time. We write the model to be used in estimation as

$$Y_i = \mu_Y + (X_i - \mu_x)\theta + e_i \ ,$$

$$\mu_x = E(X) \ ,$$

$$e_i \sim \text{Ind}(0, \sigma_e^2) \ .$$

Let $\hat{\mu}_x$ be the generalized least squares estimator of μ_x. Then our estimator for the mean of Y is

$$\hat{\mu}_Y = \bar{y} + (\hat{\mu}_x - \bar{X})\hat{\theta} \ ,$$

where $(\bar{y}, \bar{x})$ is the mean vector for the elements observed at every time period, and θ - hat is the vector of regression coefficients obtained in the regression of Y(i) on X(i) using the set of complete observations. Let m be the number of complete observations. Then the variance of the estimator is, approximately

$$V(\hat{\mu}_Y) = m^{-1}\sigma_e^2 + \theta' V(\hat{\mu}_x)\theta \ ,$$

where $V(\hat{\mu}_x)$ is the covariance matrix of $\hat{\mu}_x$.

The least squares estimator we have described will perform well in most situations. However, it is possible for the estimator to produce negative estimates for quantities known to be non-negative. This is because the estimator is linear and it is possible for some of the weights to be negative. Procedures have been developed to avoid this problem. See Huang and Fuller (1978).

4. THE U.S. NATIONAL RESOURCE INVENTORY

The Iowa State Statistical Laboratory cooperates with the U.S. Soil Conservation Service on a large survey of land use in the United States. The survey was conducted in 1958, 1967, 1975, 1977, 1982, and 1987. A survey is currently being planned for 1992.

The survey collects data on soil charcteristics, land use and land cover, potential for converting land not used for crops to cropland, soil and water erosion, and conservation practices. The data are collected by employees of the Soil Conservation Service. Iowa State University has responsibility for sample design and for estimation.

The sample is a stratified sample of the non federal area of 49 states (all except Alaska) and Puerto Rico. The sampling units are areas of land called segments. The segments vary in size from 40 acres to 640 acres. Data are collected for the entire segment on items such as urban land and water area. Detailed data on soil properties and land use are collected at a random sample of points within the segment. Generally, there are three points per segment, but 40 acre segments contain two points and the samples in two states contain one point per segment. Some data, such as total land area and area in roads, is collected on a census basis external to the sample survey.

In 1982 the sample contained about 350,000 segments and nearly one million points. The 1987 sample was composed of about 100,000 segments. The majority of the 1987 sample segments were a subsample of the 1982 segments. However, about 1500 new segments were selected in areas of rapid urban growth. Data were collected on about 280,000 points in 1987.

For the first time in 1987, it was decided that longitudinal data analysis would be performed for the period 1982-1987. Also for the first time, it was decided that the data were to be made available to the state Soil Conservation Service staff so that they could perform their own analyses.

In 1987, the field personnel were provided with a preprinted work sheet containing the 1982 information for the segment. They entered the information for 1987 on the basis of field observation and aerial photography. Field personnel were permitted to change the 1982 data if they found it to be incorrect. Edit and checking procedures were applied throughout the processing operation.

The sample was designed to produce reasonable estimates for units called Major Land Resource Areas. These areas are defined on the basis of soil and cover characteristics. There are about 180 Major Land Resources Areas in the study area. Also the acreage estimates for any county were to be consistent with the total acreage of that county. There are about 3100 counties in the sample. Because the sample must provide consistent acreage estimates for both counties and Major Land Resource Areas, the basic tabulation unit is the portion of a Major Land Resource Area within the county. There are 5530 of these units, which we called MLRAC's.

The design of the sample is a simple form of a panel survey in that the 1987 sample is nearly a subsample of the 1982 sample. It was decided to use as the control variables from the 1982 study, the 1982 acres of 14 major land uses such as cropland, rangeland, forestland, and urban land. In addition, the external information, such as 1987 area in roads, and the segment information, such as 1987 area in urban land, is auxiliary information similar to that obtained from incomplete observations.

Table 4 is a condensed version of an estimation table for one of the states in the survey. It contains only four uses instead of the 14 actually employed in the estimation. The entries in the right column are the 1982 estimates. The entries in the last row for urban land and roads are from the segment data and the external sources, respectively. The vector of six entries, (the first four entries of the last column, 1987 urban land, and 1987 roads) is a vector of totals corresponding to the vector of estimated means, $\mu(x)$ - hat of Section 3.

The internal estimates of the table are essentially least squares estimates that satisfy the six control totals. In the actual estimation scheme it was necessary to use imputation methods when, for example, a change is reported in the segment data, but there is no corresponding change in the point data.

Table 4. Illustration of estimation procedure

	1987				
1982	Cropland	Other	Urban	Roads	TOTAL
Cropland	26,243	179	13	6	26,441
Other	771	7,114	6	2	7,893
Urban	0	0	623	0	623
Roads	17	4	0	1,038	1,059
1987 TOTAL	27,031	7,297	642	1,046	36,016

The design produced large variances for the directly estimated change in small uses such as urban land, farmsteads, and small water bodies. Therefore, a small area estimation scheme was used to construct estimates of change for the major land resource areas within counties. We used a computer program for small area estimation that we have developed at Iowa State University. The theory for the small area estimation procedure is decribed in

Fuller (1986). Estimated changes in five small land uses for each of the 5,500 MLRAC's were constructed with the small area program. This procedure is essentially an allocation program in that the sum of the MLRAC estimates is the state estimate. Estimates for the entries in Table 4 (with 14 categories) were constructed for each MLRAC. In this estimation, the small area MLRAC estimates, the external estimate for roads, and the state marginals for cropland were used as controls. The final step in the estimation procedure was the assignment of weights to the point data such that the weighted point data give the estimates of Table 4 for each MLRAC.

To summarize, the final product of the estimation procedure is a tabulation data set of points that permits estimation of complete two-way tables of 1982-1987 land use for any identifiable area designation. The estimates are consistent with previous estimates for major land use categories for the states and are consistent with data from sources outside of the point sample.

Generally speaking, it is not possible to obtain good variance estimates from the tabulation sample, although segment and stratum identification are given in the data set. Variance estimates computed with the point data for principal uses, such as cropland, will be too large because of the control on the larger 1982 sample.

5. MEASUREMENT ERROR

Measurement error can have a very large impact on the analysis of data over time. This impact may be moderate in the case of simple means reported at a sequence of times. However, in gross change estimation and in regression estimation, measurement error can be extremely important.

To illustrate the magnitude of measurement error bias in estimators of gross change, let us return to the simple example of Table 1. If the data were collected by a procedure such as that of the U.S. Census Bureau, the work of Chua and Fuller (1987) demonstrates that the interior cells of the two way table will be seriously biased. Also see Abowd and Zellner (1985) and Poterba and Summers (1985). Under the Chua-Fuller model, the response error at the two points in time is assumed to be independent. Also it is assumed that, at each time,

$$P(\text{response} = E \,|\, \text{true} = E) = 1 - \alpha + \alpha P_E \ ,$$

$$P(\text{response} = U \,|\, \text{true} = E) = \alpha P_U \ ,$$

$$P(\text{response} = U \,|\, \text{true} = U) = 1 - \alpha + \alpha P_U \ ,$$

$$P(\text{response} = E \,|\, \text{true} = U) = \alpha P_E \ ,$$

where α is the parameter of the response mechanism. Under this model the expected value for the proportion unemployed at any point in time is the true proportion. A consistent estimator of P(EE) under the Chua-Fuller model is

$$\hat{\pi}_{EE} = (1 - \alpha)^{-2}(\hat{P}_{EE} - \hat{P}_{E.}\hat{P}_{.E}[1 - (1 - \alpha)^2]) \ ,$$

where $\hat{P}(EE)$, $\hat{P}(E.)$ and $\hat{P}(.E)$ are the direct estimators and α is a parameter of the response mechanism. Also see Battese and Fuller (1973). On the basis of the U.S. reinterview data, a value of $\alpha = 0.10$ is not unreasonable. For our example, we have

$$\pi_{EE} = (0.90)^{-2}(0.91 - 0.93(0.94)(0.19))$$

$$= 0.9184 \ .$$

The corresponding two-way table of proportions adjusted for response error is

$$\begin{pmatrix} 0.9184 & 0.0116 \\ 0.0216 & 0.0484 \end{pmatrix} .$$

In this example, the bias in the direct estimator of $P(EE)$ is 0.0084. Chua and Fuller estimate the bias to be about 0.0168 in the three way table that includes the not-in-the-labor-force category. Table 5 contains a comparison of alternative estimation procedures for $P(EE)$. A sample of 10,000 is assumed. The first three procedures are those of Table 3. The last three are the three estimators adjusted for measurement error bias. In the variance calculations, α is assumed to have a standard error of 0.01. The estimators of $P(E.)$ and $P(.E)$ are not changed by the adjustment for measurement error bias. In this example the squared bias in the ordinary estimator of $P(EE)$ is about nine times the variance of the generalized least squares estimator. Thus the measurement error bias dominates the mean square error of the estimator of $P(EE)$.

Table 5. Mean square error of alternative estimators for a sample of 10,000 at each time and 50% overlap (Mean square error of measurement error adjusted GLS = 100.)

	Procedure					
	Ordinary			Measurement Error		
Parameter	Simple	Rest. GLS	Full GLS	Simple	Rest. GLS	Full GLS
$P_{E.}$	111	111	100	111	111	100
$P_{.E}$	111	101	100	111	101	100
P_{EE}	1071	967	961	250	106	100

These results have serious implications for survey design. To illustrate this, we return to the gross change problem. Assume that our objective is to estimate the probability that a person will remain employed for two periods, $P(EE)$. We assume that it is possible to conduct independent reinterviews for each point in time, and that interviews at two points in time are independent. We assume that the only interview procedures permitted are:

A. Interview and reinterview at one of the times.
B. Interview at time one and interview at time two.

We assume that the response error is unbiased and that a simple two-class (employed and unemployed) model is appropriate. We also assume that the probabilities of correct response depend only on the current class of the respondent.

Let the response probabilities be defined in terms of α and let

$$\gamma = (1 - \alpha)^{-2} .$$

Let $\theta(ij)$ denote the ij-th element of the 2×2 matrix of probabilities observed in the reinterview study. That is, $\theta(ij)$ is the probability that an individual responds i on the first interview and j on the reinterview. For this simple model we can obtain explicit expressions for the estimators. We have

$$\hat{\gamma} = (\hat{\theta}_{11} - \hat{\theta}_1^2)^{-1} (\hat{\theta}_1 - \hat{\theta}_1^2)$$

and

$$\hat{P}_{11} = \hat{\gamma} (\tilde{P}_{11} - \tilde{P}_{1.} \tilde{P}_{.1}) + \tilde{P}_{1.} \tilde{P}_{.1} ,$$

where

$$\theta_1 = \theta_{11} + \theta_{12} = \theta_{11} + \theta_{21} ,$$

$\hat{\theta}(ij)$, are the estimates from the reinterview study and $\tilde{P}(ij)$ are the estimates from the interviews conducted at the two time periods.

In constructing the estimator, the reinterview study is used only to estimate the measurement error parameter. In fact, the reinterview study could be used in a generalized least squares procedure to improve the estimates of $P(11)$, $P(1.)$, and $P(.1)$. Under the assumption that all interviews are of equal cost, it can be demonstrated that about one fourth of the resources should be used for the reinterview study. The relative efficiency of the measurement error procedure to the direct biased procedure is given in Table 6.

Table 6. MSE efficiency of MEM to direct

	Sample size, n			
	500	1,000	5,000	10,000
MSE direct/MSE MEM	0.87	1.13	3.22	5.84

In small samples the direct procedure has a smaller mean square error because of the smaller variance. Recall that only three fourths of the observations furnish information on $P(EE) = P(11)$. However, for samples greater than 750, the squared bias dominates the mean square error of the direct procedure and the consistent measurement error procedure has a smaller mean square error. This small example demonstrates the efficacy of surveys containing a component to estimate the parameters of the measurement process.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 68-3A75-8-12 with the Soil Conservation Service, U.S. Department of Agriculture. I thank Margot Tollefson for the computations.

REFERENCES

Abowd, J. M. and Zellner, A. (1985), "Estimating Gross Labor Force Flows," Journal of Business and Economic Statistics, 3, 254-283.

Battese, G. E., Hasabelnaby, N. A., and Fuller, W. A. (1989), "Estimation of Livestock Inventories Using Several Area - and Multiple-Frame Estimators," Survey Methodology, 15, 13-27.

Chua, T. C. and Fuller, W. A. (1987), "A Model for Multinomial Response Error Applied to Labor Flows," Journal of the American Statistical Association, 82, 46-51.

Cocharan, W. G. (1942), "Sampling Theory When the Sampling Units are of Unequal Sizes," Journal of the American Statistical Association, 37, 199-212.

Duncan, G. J. and Kalton, G. (1987), "Issues of Design and Analysis of Surveys Across Time," International Statistical Review, 55, 97-117.

Eckler, A. R. (1955), "Rotation Sampling," The Annals of Mathematical Statistics, 26, 664-685.

Garcia, P. A., Battese, G. E., and Brewer, W. D. (1975), "Longitudinal Study of Age and Cohort Influences on Dietary Patterns," Journal of Gerontology, 30, 349-356.

Graham, J. E. (1973), "Composite Estimation in Two Cycle Rotation Sampling Designs," Communications in Statistics, 1, 419-431.

Gurney, M. and Daly, J. F. (1965), "A Multivariate Approach to Estimation in Periodic Sample Surveys," Proceedings of the Social Statistics Section of the American Statistical Association, 242-257.

Huang, E. T. and Fuller, W. A. (1978), "Nonnegative Regression Estimation for Sample Survey Data," Proceedings of the Social Statistics Section of the American Statistical Association, 300-303.

Jessen, R. J. (1942), "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," Iowa Agricultural Experiment Station Research Bulletin, 304, 54-59.

Jones, R. G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," Journal of the Royal Statistical Society, Ser. B, 42, 221-226.

Kalton, G. (1983), Compensating for Missing Survey Data, University of Michigan, Survey Research Center.

Kasprzyk, D., Duncan, G. J., Kalton, G., and Singh, M. P. (1989), Panel Surveys, New York: John Wiley.

Kasprzyk, D. and McMillen, D. B. (1987), "SIPP: Characteristics of the 1984 Panel," Proceedings of the Social Statistics Section of the American Statistical Association, 181-186.

Lazarsfeld, P. F. and Fiske, M. (1938), "The Panel as a New Tool for Measuring Opinion," Public Opinion Quarterly, 2, 596-612.

Lepkowski, J. M. (1989), "Treatment of Wave Nonresponse in Panel Surveys," in Panel Surveys, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: John Wiley.

Little, R. J. A. and Rubin, D. B. (1987), Statistical Analysis with Missing Data, New York: John Wiley.

Little, R. J. A. and Su, H. L. (1989), Item Nonresponse in Panel Surveys, in Panel Surveys, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: John Wiley.

Madow, W. G., Olkin, I., Nisselson, H., and Rubin, D. B. (1983), Incomplete Data in Sample Surveys. (Three volumes) New York: Academic Press.

Patterson, H. D. (1950), "Sampling on Successive Occasions with Partial Replacement of Units," Journal of the Royal Statistical Soc., B12, 241-255.

Poterba, J. M. and Summers, L. H. (1985), "Adjusting the Gross Change Data: Implications for Labor Market Dynamics," Proceedings of the Conference on Gross Flows in Labor Force Statistics, U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, pp. 81-95.

Rao, J. N. K. and Graham, J. E. (1964), "Rotation Designs for Sampling on Repeated Occasions," Journal of the American Statistical Association, 59, 492-509.

Smith, T. M. F. and Holt, D. (1989), "Some Inferential Problems in the Analysis of Surveys Over Time," paper presented at the 47th session of the International Statistical Institute, Paris.

Wolter, K. (1979), "Composite Estimation in Finite Populations," Journal of the American Statistical Association, 74, 604-613.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

UNIQUE FEATURES AND PROBLEMS OF ROLLING SAMPLES

L. Kish[1]

I am grateful for this opportunity to explain the chief features of rolling samples and the purposes they are meant to serve. First, let me attempt a definition of rolling samples: A combined (joint) design of F separate periodic samples, each a probability sample of the entire population, designed so that the cumulation of the F periods yields a detailed census of the whole population; also intermediate cumulations should yield details intermediate between 1 and F periods. We may appreciate that definition by looking at examples and counter-examples. We shall also examine possible variations that would satisfy the definition and the conflicting needs that rolling samples can be aimed to meet.

Imagine a weekly national sample, each with epsem selection rates of 1/520, and so designed that in 520 weeks they are "rolled over" the entire population and the cumulation yields a complete census of the population averaged over ten years. Each year would yield national and local samples with selection rates of 52/520 = 1/10. The design would combine weekly national samples with an averaged decennial complete census, and with sample censuses of ten percent each year.

I use the words "would" and "might", because the design does not yet exist anywhere as far as I know. I am bold to coin this definition, as I first used it in 1981 in a published report to a committee of the U.S. Congress [Kish 1981]. Earlier I described such plans with the title "Rotating samples instead of censuses" [Kish 1979]. But the name "rotating samples" met objections, because of its confusion with the well known partially overlapping samples that are widely used for labor surveys. By coining the new name "rolling samples" I intend to avoid needless confusions with other designs. By coining descriptive names for my methods I also hope to advance understanding. Furthermore, they also help to avoid having the authors' names attached to their methods; an annoying practice, which leads to needless antagonism about priorities.

The labor force surveys now in use, such as the CPS in the USA and the CLFS in Canada, differ from rolling samples in important ways. First, the labor force surveys typically have considerable overlaps, which hinder and delay cumulations. Second, they are confined to primary sampling areas, so that cumulations fail to cover the national population area. Third, they may not be large enough in size to cumulate to a complete census. Fourth, the methods tend to yield less complete coverage than the census.

Nonoverlapping samples are sometimes called "multiround" surveys and are used to cumulate data that depend on short periods of recall. In developing countries they have been used to collect demographic data, such as birth and death rates. The 52 weekly nonoverlapping samples of 1000 households of the HIS of the NCHS may be a good example. However it is also too small and too confined to PSU's to qualify now as a rolling sample, which would yield a detailed national census.

At the August 1989 meeting of the ASA, a statewise alternation between years of complete censuses was advanced, with the false tag of "rolling samples". I asked the authors to avoid this needless confusion. Furthermore, I hope that the idea has little chance of success. It would confound yearly and statewide variation, so as to confuse both temporal and spatial comparisons. As such it would contradict the efforts of the UN for decennial collection dates. We may forego further criticisms here, but only use it as a counterexample to rolling samples.

We now must review, ever so briefly, the chief alternatives to rolling samples for providing the detailed information needed for small domains, which is one principal aim of cumulating rolling samples, the other being to provide overall population estimates at frequent (weekly, yearly) intervals. Publicity today favors detailed population counts for administrative areas, but details for other domains, for "cross domains" (like age and social classes) may be as important in the long run.

First and foremost we must put the decennial censuses of population, housing, agriculture, industry and others, which humankind has been spreading over the earth's face in the last two centuries, and especially in the last two generations with the help of the United Nations. In addition to the detailed data for small domains, censuses sometimes may also obtain better coverage due to concentrated publicity and the national "ceremony" connected with the census. The Chinese census of 1982 is a good example. The concentrated efforts of the census may also yield lower unit costs than surveys; but at 2.6 billions, the US census of 1990 will cost $10 per capita or $30 per household. However,rolling samples are being proposed here chiefly because decennial censuses lack timeliness: from collection to use the census data are typically from about 1 to 14 years old [Kish 1981].

More frequent censuses, quinquennial or yearly, have also been proposed. But quinquennial may not be frequent

[1] L. Kish, Research Scientist & Prof Emeritus, Inst. for Social Research, University of Michigan, Ann Arbor, MI 48106.

enough, and yearly censuses would be too costly. Sample censuses of 1 or 10 percent have been proposed, but the former may be too small and the latter too costly. In two countries at least, quinquennial censuses of 10 percent had half of the cost of a complete census and one also suffered from increased noncoverage. The 1 percent microcensus of West Germany and the 1/2000 samples of China provide some yearly data. Canada had 10 percent census in 1985. I doubt that these efforts will provide generally the needs for data that are both timely and detailed. To paraphrase Lincoln: "You cannot poll all the people, all of the time".

That stolen phrase leads us to registers or administrative records as a method for collecting data, which can be both timely and detailed. Outstanding examples are the population registers of the nordic countries: Sweden, Norway, Denmark, and Finland, and perhaps a few others. In a few cases they have replaced or may replace censuses with data from registers. Their completeness is based on cooperation, motivation, and literacy. In other situations their coverage, quality, and updating are far from adequate. I expect further growth in their quality, their spread and their use, but not that such registers will replace censuses either soom or completely, because their contents are likely to be limited to a few basic variables, too few for modern census needs. To paraphrase Lincoln again: "You cannot poll all the people, all the time, about everything".

What about synthetic and raking estimators that would give us timely and detailed estimates based on censuses, plus registers, plus surveys? I am optimistic about progress with those methods, but not about their replacing data collections by censuses or by rolling samples.

Now we must discuss briefly three problems facing cumulated rolling samples: their costs, their coverage, and their bases in averaging over changing populations. These averages must cover both population changes over time and individual changes of location in space.

Averaging variations over time must overcome mental blocks based on the tradition and practice for both censuses and surveys. I have made several efforts to overcome those blocks with arguments based on statistical inference and philosophy, and we need more theoretical, methodological, and empirical work. The sum of repeated surveys over an enture time interval can lead to better statistical inference than a single, concentrated one-shot survey. Probability selection of time segments from an entire interval permits statistical inference from the sample to an average condition over the interval. On the contrary, inference from a "typical" time segment from one-shot survey to the entire interval demands judgment, assumptions, models about the nature of variation, or lack of variation, over the entire interval. The choice of a single time segment is exposed to the risks of seasonal, cyclical, secular, and catastrophic variations, known or unknown, The sum of repeated surveys relies on averaging out the variations over the repeated surveys [Kish, 1965, 12.5D]. Sampling and cumulating over time should be preferable on statistical, methodological grounds to accepting any arbitrarily chosen "typical" period. It is paradoxical that judgmental selection is still accepted and practiced in the time dimension, whilst we refuse to tolerate judgmental selection of spatial segments in probability sampling [Kish, 1979, 1981, 1983, 1986].

A less formidable but annoying problem for rolling samples is caused by changes of location (of people, households etc) so that the same units can fall into two or even more periodic samples. These moves are ruled out by the arbitrary census date, though their application is costly, arbitrary and faulty. They also occur in one-shot surveys. But they will occur by the thousands in cumulated rolling samples. However for the random selections of area segments of probability samples they cause no bias. We only need to understand and explain.

The problems of cost for a complete rolling census seem formidable compared to the costs of most periodic surveys alone. But the contrasts are less formidable in smaller countries, because sampling fractions are greater in smaller countries. For example, monthly labor force surveys of 80,000 households need only f=1/1000 in a giant country of 80 million households; but they need f=1/100 in a country of 8 million households, and those would cumulate without overlaps to a complete census in the 120 months of ten years. We shall discuss the overlaps soon. Furthermore the cost per interview for a rolling sample is bound to be higher than for the current samples confined to primary sampling areas. However the travel cost would not increase nearly as much as would be suggested by small areas of PSU's on the maps of the national territory. The large majority of the sample and the population in each country are concentrated in a relatively small number of "self-representing areas".

However, for allowable cost we must add to the costs of labor force surveys, also the costs of the decennial and perhaps quinquennial censuses, because the rolling samples presume to do the work of both. It may be true that census workers are generally poorly paid but the costs of recruiting and training for only a few days work may be relatively high.

The question of relatively good coverage by some censuses compared with sample surveys, as mentioned earlier, is too technical and specific for brief treatment here. It is likely that with special efforts, the coverage in sample surveys may be improved. For example, the USCB is endeavoring to check and improve the 1990 Census with a special sample survey of 150,000 households.

For periodic labor force surveys, and for some others also, considerable overlaps are often used for two chief reasons. The most important reason is less often mentioned: the later interviews cost less than the first, especially when they are done by telephone. Though the ratios of their total costs are not overwhelming, they demand consideration in any comparison. Better known in formulas are the reasons based on the correlations found in reinterviews; these reduce variances modestly for current estimates, and even more for estimating net (or macro-) changes between periods. However those correlations are weak for many survey variables, for example for measures of unemployment. They are further weakened by response errors and by moving rates that approach 0.2 between years.

Thus correlations are lower when the overlaps are for segments rather than persons; but such overlaps are simpler to handle, are cheaper, and not subject to the biases of panels of persons. On the other hand, a panel of persons would have higher correlations and also permit the analysis of individual changes, i.e., micro changes, or gross changes. Because of this conflict some surveys have done both: covered the same segments and also followed the moving individuals for panels.

The size and nature of the overlapping sample needs technical studies; these studies should be multipurpose because the correlations will vary greatly between variables. My informal advice is for overlaps that would be 1/3 or less of the cumulated nonoverlapping portion. Also the overlap could be a panel of individuals followed for many periods, to permit dynamic analysis of individual changes, now missing from labor force surveys. The overlaps over many periods would reduce variances of net changes for all those pairs of periods. I proposed the name split panel designs (SPD) for such designs [Kish 1982, 1986, 1987].

The basic design calls for F periodic surveys for frequent (weekly or monthly or yearly) population estimates designed for cumulating the F samples for small domain estimates over the entire interval. Within that definition a great deal of flexibility may be encouraged, and some examples now follow. First, improved estimates for domains (provinces) may be designed both with larger sampling fractions and with longer cumulations. With quarterly estimates instead of monthly, and tripled sampling rates, the sample base would be increased by a factor of nine, for example.

Periodic symmetrical samples (weekly or monthly or quarterly) may be the simplest and best, but departures from that may be tolerated, and perhaps compensated with weights. Furthermore, to the basic contents of the surveys, additional variables may be added as needed.

Although the emphasis has been on the two extremes — single surveys for timeliness and complete cumulations over the entire interval (ten years?) for small domains — intermediate cumulations for major domains (provinces?) and for minor domains (districts?) would be often desirable and feasible. At this point we must add that whereas a complete 100 percent census was indicated or implied, the basic idea can also include large fractions (10 percent) as the census targeted over the interval; particularly where decennial censuses are also collected. Similarly the "population" in the definition is clearly meant to include many populations along with a national count of persons.

The sizes and weights of periodic samples should be considered together, and a great deal of flexibility is advisable. Methodological research can make solid contributions. Here and now, we assume similar sampling fractions for all periods, and only consider different weights for each of 10 years, with a total weight of 10 over the entire 10 years. For the national sample and for highly fluctuating variables (e.g. infectious diseases) the last year may carry the full weight of 10. On the contrary, for total populations of small domains, each of the ten years may have a unit weight of 1. However many variables and for large domains an intermediate moving average may be better than either extreme, for example (100, 90, 80, 65, 50, 40, 30, 20, 15, 10)/50.

Finally, I wish to add a quote (Kish, 1986) that concerns rolling samples as well as other periodic (or other repeated) sample surveys. "Fifth, statistical strategy should dictate less frequent reporting especially for small but non-negligible domains. Too often such domains either remain unreported, or they are reported with unduly large errors or at too great a cost, or both. As prime examples consider the vast but underpopulated areas with small populations which appear in many countries and for which provincial authorities demand separate reports. Other examples come from demographic, ethnic, occupation groups, etc., for which separate data are needed. Instead of the usual rigid practice prevailing now, it would be preferable to report at, to cumulate for, and to design for longer periods for these smaller domains. The tables for these statistics should indicate the different designs used for those statistics."

REFERENCES

Kish, L. 1965. *Survey Sampling*, New York and Sydney: John Wiley and Sons.

Kish, L. 1979. Rotating Samples Instead of Censuses, *Census Forum*, Honolulu: East-West Center 6, 1-13.

Kish, L. 1981. *Using Cumulated Rolling Samples*, Washington: Library of Congress, U.S. Government Printing Office, No. 80-528.

Kish, L. 1983. Data Collection over Time and Space, in T. Wright, *Statistical Methods and the Improvement of Data Quality*, Orlando: Academic Press, 73-84.

Kish, L. 1986. Timing of Surveys for Public Policy, *Australian Journal of Statistics*, 28(1), 1-12.

Kish, L., and Verma, V. 1986. Complete Censuses and Samples, *Journal of Official Statistics*, 2, 381-94.

Kish, L. 1987. *Statistical Design for Research*, New York: John Wiley and Sons, Section 6.5.

Kish, L. 1989. *Sampling Methods for Agricultural Surveys*, Rome: FAO, Statistics Division, Section 16.3.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

SAMPLE MAINTENANCE BASED ON PEANO KEYS

Kirk M. Wolter and Rachel M. Harter

ABSTRACT

We discuss frame and sample maintenance issues that arise in recurring surveys. A new system is described that meets four objectives. Through time, it maintains (1) the geographical balance of a sample; (2) the sample size; (3) the unbiased character of estimators; and (4) the lack of distortion in estimated trends. The system is based upon the Peano key, which creates a fractal, space-filling curve. An example of the new system is presented using a national survey of establishments in the United States conducted by the A. C. Nielsen Company.

1. INTRODUCTION

We are concerned with recurring surveys conducted over time and the maintenance they require. Let $\mathcal{U}_t$ denote a survey universe at time t, with t = 0 denoting the inception of a new survey. We assume a probability sample of units of $\mathcal{U}_0$ has been selected, and thus that it is feasible to construct unbiased (or at least consistent) estimators of the population total and other parameters of interest. As time goes by, we assume the universe is surveyed repeatedly at regular intervals of time, in part to track the "level" of the population, and in part to measure its "trends." A panel or a rotation sampling design is usually employed for this purpose (see, e.g., Rao and Graham (1964) and Wolter (1979) and the references cited by those authors). In all such surveys of people or their institutions, which is all we concern ourselves with here, the composition of the universe changes with time as births, deaths, and other changes occur to the status of the units. The survey frame, the sampling design, and the schemes for observing or collecting the survey data must be maintained for such change; otherwise, the sample may become excessively biased and cease to be representative of the universe.

The types of maintenance issues that arise in recurring surveys depend in part on the kind of universe under study, in part on the choice of sampling unit, and in part on the interplay between the sampling unit and the universe elemental units. We shall summarize briefly the issues that arise in four different situations:

(i) establishment surveys with establishment as the sampling unit;

(ii) establishment surveys with company or some similar cluster of establishments as the sampling unit;

(iii) surveys of people or households with the address or housing unit as the sampling unit; and

(iv) surveys of people or households with the household or family as the sampling unit.

In this work, we use the words "establishment" and "company" in a generic sense. An establishment may be a retail store, a manufacturing plant, a school, a hospital, a golf course, or any other similar, single-location entity, while the corresponding company would be the corporate, legal entity that owns the retail store, or the school district, and so on. In some cases, of course, the establishment and company will be synonymous, e.g., a single, independent grocery store.

For case (i), the main universe dynamics include

- establishments arising from new construction
- reclassified establishments from some out-of-scope category to an in-scope category
- reclassified establishments from one in-scope category to another in-scope category
- reclassified establishments from an in-scope category to an out-of-scope category
- conversion of a structure from residential use to commercial use
- conversion of a structure from commercial use to residential use

[1] Kirk M. Wolter, Vice President, A.C. Nielsen, Northbrook, Illinois, 60062
[2] Rachel M. Harter, A.C. Nielsen, Northbrook, Illinois, 60062

- demolition of an existing establishment

- establishment that moves in and out of vacancy status

- changes in the configuration of an establishment, e.g., division into two or more establishments.

Case (ii) is far more complicated than case (i), principally because sampling units are now clusters of elemental units. All of the issues from case (i) apply to single-establishment companies. For multi-establishment companies, we face the following additional dynamics:

- mergers wherein two companies combine to form a new successor company

- acquisitions wherein one company is acquired by another, with the acquiring company as the sole successor company

- joint ventures wherein two companies collaborate to form a new company that may be a subsidiary to both the parent companies

- divestitures wherein a company spins off a new and independent company

- divestitures where a company sells parts of itself to another acquiring company.

In a sense, case (iii) is very similar to case (i) in respect to the kinds of universe dynamics that may arise:

- housing units arising from new construction

- reclassified housing units from some out-of-scope category to an in-scope category

- reclassified housing units from one in-scope category to another

- reclassified housing units from an in-scope category to an out-of-scope category

- conversions from residential to commercial

- conversions from commercial to residential

- demolition of an existing housing unit

- reconfigurations of existing structures, e.g., reconfigurations of apartments within a small multiunit structure.

Note how closely these issues match those for case (i).

Finally, case (iv) is very similar to case (ii) in terms of the composition and complexity of universe change. Maintenance issues include:

- marriage, wherein a new successor family is created, possibly from whole predecessor families or from part families

- new members move into an existing family, either eliminating another family or part of a family

- divorce, wherein successor families may be created from one predecessor family

- family members move away, either to join another existing family or to establish a new family

- births of family members

- deaths of family members

- a whole family moves, thus requiring tracing and perhaps altering field-work assignments.

To handle the universe dynamics listed above, properly reflecting them in the sample, so that sample representativeness is retained over time, the survey organization must design and adopt an explicit system of maintenance. We define a sample maintenance system to be a sampling design and a universe updating methodology, possibly specified in the form of simple rules, that permit the statistician to achieve known, nonzero probabilities of inclusion for each of the elemental units in the population for each time period in the recurring survey, or failing that, to weight the survey data properly so as to achieve unbiased or consistent estimators of the population parameters of interest. From cases (i) through (iv) above, it is clear that a maintenance system must perform at least four functions:

- give new elemental units a known, nonzero probability of selection
- account properly for elemental units that may no longer exist in a substantive sense
- not give elemental units multiple chances of selection into the sample; otherwise, if multiple changes are given, the system must appropriately record this information so that adjustments may be made in the estimation procedures
- appropriately update the universe frame so as to facilitate and control the above activities.

A general and necessary rule of thumb for any sample maintenance system is that the system, or the rules that define the system, must treat symmetrically universe changes both within and outside of the sample. If a proposed maintenance rule violates this rule of thumb, then there is risk of bias in estimators of totals and other universe parameters to be estimated. For example, consider two rules that might be used for case (ii) for sampling new companies created as the result of a divestiture. One possibility is to declare the new companies part of the sample _if_ their predecessor companies were part of the sample, and otherwise, if their predecessors were not part of the sample, to subject the new companies to a new round of sampling. This rule is seen to give the new companies multiple probabilities of selection, and thus may result in biased estimation unless appropriate adjustments are made in the estimation procedure. (The adjustments we have in mind are related to the multiplicity rules studied by Monroe Sirken (1970) and others.) A second possibility is to declare the new companies part of the sample _if and only if_ their predecessor companies were part of the sample. Because this second rule treats symmetrically the universe changes both within and outside of the sample, it is seen to result in unbiased estimation for the survey parameters of interest.

In designing a sample maintenance system, the statistician must be guided not only by the statistical properties of the resulting estimators, but also by the cost, feasibility, and customer acceptance of alternative rules. Some rules may require additional data collection, thus entailing additional cost that must be planned from the inception of a new recurring survey. Certain applications may actually require that additional data be collected retrospectively. This may be impractical, or at the very least, may entail considerable nonsampling error, thus risking bias. Some rules may well be feasible and cost-effective, yet may not satisfy the requirements of the customers or users of the survey data.

Finally, we note that this problem of maintenance is neither new nor newly recognized; for example, maintenance systems have been in place for years in many of the major recurring surveys at Statistics Canada, the United States Bureau of the Census, and the A. C. Nielsen Company. Nevertheless, there is remarkably little literature on this subject. For brief discussions of some maintenance issues, see Wolter et al (1976) for case (ii), Hanson (1978) for case (iii), and Ernst (1989) for case (iv).

In the balance of this article, we focus on case (i), where the establishment is both the sampling and elemental unit. This is the case we face in our establishment surveys at the A. C. Nielsen Company. Section 2 describes one of our major surveys, the Scantrack survey, and the specific maintenance issues we face in that survey. We also describe some of the key objectives we had in designing a new maintenance system for this survey.

The new maintenance system is based upon a parameter known in mathematics as the Peano key, which creates a fractal, space-filling curve. The Peano key is defined in Section 3, where we also provide several graphical displays for illustration purposes. We close the article in Section 4 by describing the rules that implement our new maintenance system.

2. THE SCANTRACK SURVEY

The Nielsen companies provide information from several marketing surveys. The media surveys, such as Nielsen Television Index and Nielsen Station Index, are based on samples of either housing units or households. Surveys for the packaged goods industry, including Nielsen Food Index, Nielsen Drug Index, and Nielsen Scantrack United States (NSUS), are based on samples of stores. The Single Source service, which ties together consumer purchasing behavior with household television viewing and retail marketing support, is based on both household and store samples. Although sample maintenance is an important issue to each of these surveys, the present discussion will focus on our Scantrack sample of grocery supermarkets which is the basis for the NSUS service.

The Scantrack sample includes 3,000 supermarkets, stratified by 50 metropolitan markets and a remaining United States stratum. Within a market, the sample is further stratified by major chain organizations. The frame is ordered geographically, and a systematic sample is selected within each stratum to achieve proper socio-economic representation. This sample is also representative of store age, store size, and other factors associated with item sales. Although a geographically ordered systematic sample is exceedingly simple and straightforward, the choice of this sample design is justified based on years of experience, as well as the results of empirical studies in which various sample designs were tested on universe data.

Stores in the Scantrack sample are equipped with electronic scanners at the checkout, which read bar codes on packaged goods. Bar codes are called universal product codes or UPC's. When the item is scanned, the transaction is entered into the store's computer where the UPC is matched with the item's price. Each week, the sample stores provide us with total sales movement and price data for every item that is scanned in the store. Since a supermarket typically carries over 10,000 UPC's, we receive and process over 30 million observations per week.

In addition to scanner data, we obtain data on promotion conditions for the items in each of the sample stores, including whether an item was featured in a newspaper advertisement, store display, or store coupon. If an item was featured, we also know the type of newspaper advertisement used and the location of the display within the store.

NSUS reports include estimated sales totals for individual items and aggregates of items for each market and the total United States. A ratio estimator is used, with all-commodity volume as the auxiliary variable. All-commodity volume, or ACV, refers to total sales of all items in a store, usually on an annual basis. ACV tends to be highly correlated with sales of individual items. In addition, the NSUS reports include estimates of sales and sales rates by promotion condition and estimates of year-to-year sales trends.

Continuous maintenance is necessary for the Scantrack sample because the national supermarket universe of approximately 30,500 stores is not static. In a recent 12-month period, approximately 2,200 new supermarkets opened, and 2,450 existing stores went out of business. Another 170 stores were reclassified during the year. Reclassification can result from any of a number of changes. Some smaller grocery stores enter the Scantrack universe when their ACV's surpass the $2-million-per-year threshold which defines a supermarket. A store might change name or location, or be expanded through remodeling. Some stores change to an extended or economy format, such as a superstore, warehouse store, or other nontraditional supermarket. In 1979, about 3,800 extended and economy stores accounted for 17% of total supermarket sales. By 1988, the number of extended and economy stores had grown to over 9,000, and they accounted for almost 50% of all supermarket sales (Progressive Grocer 1989). Sometimes, individual stores or entire chains are acquired by another organization, affecting stratum definitions.

In addition to universe changes, missing or faulty data situations arise that require substitution of sample stores. Some selected sample stores do not scan, and some that do have incompatible scanning equipment. If a store is consistently unable to provide us with usable data, it must be dropped from the sample. Sometimes a request for a sample change within an organization comes from the chain itself. Occasionally, a retailer simply refuses to cooperate.

The principal objectives of our maintenance system for the Scantrack sample are: (1) the sample should maintain geographic balance through time, (2) the system should maintain the sample size through time, (3) the sample should adhere to principles of probability sampling so as to avoid bias in estimators of total sales, and (4) sample changes should not disturb excessively estimates of year-to-year trends.

Geographic balance is a proxy for socio-economic balance. Because different neighborhoods have different purchasing patterns, geographical balance is important to achieving an efficient sample design (i.e., low sampling variability) over a wide range of products. Furthermore, geographic balance is an important factor in our customers' perception of an appropriate sample.

A sample size decrease would adversely affect the standard errors of the estimators, and a sample size increase would adversely affect our costs. Neither outcome is desirable. Furthermore, contracts with chain organizations specify sample sizes and cooperation payments, and any changes would have to be renegotiated. This too is undesirable.

All applications involving Scantrack data require efficient, unbiased estimators of total sales. Manufacturers and retailers need such data for everyday business decisions, such as how much to produce, how much to ship, how much to keep in inventory, and how to allocate store shelf space.

Clients also require reliable year-to-year trend information for managing their businesses. Trend estimates help manufacturers assess the overall health of their businesses. Both manufacturers and retailers benefit from knowing the longer-term performances of all major brands in all product categories.

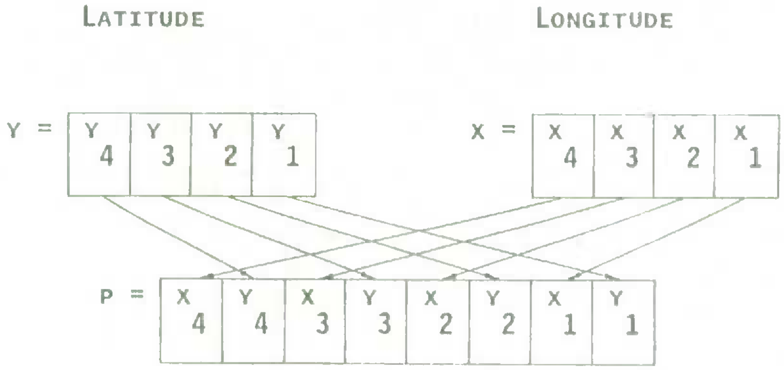
We describe the maintenance system that has been developed to meet these objectives in section 4. But first, we describe a new geographic ordering scheme in section 3.

3. PEANO KEYS

The Peano key is a parameter that defines a certain fractal, space-filling curve. It provides a mapping from $\mathbb{R}^2$ to $\mathbb{R}^1$ such that points in $\mathbb{R}^2$ or spatial objects can be arranged in a unique order (Peano order) on a list. In the application we have in mind, the spatial objects are sampling units, and the space $\mathbb{R}^2$ is represented by earth's geographic coordinate system.

We obtain the Peano key by interleaving bits. See Peano (1908), Laurini (1987), and Saalfeld, Fifield, Broome, and Meixler (1988). Let $X = X_k \ldots X_3X_2X_1$ and $Y = Y_k \ldots Y_3Y_2Y_1$ represent the longitude and latitude of an arbitrary point in k-digit binary form. Then, the corresponding Peano key is $P = X_kY_k \ldots X_3Y_3X_2Y_2X_1Y_1$. Also see figure 1 for an example for the case $k = 4$. Note how simple it is to calculate the value of P.

FIGURE 1. CREATING THE PEANO KEY BY BIT INTERLEAVING



Given k-digit (for any finite k) latitude and longitude coordinates, the spatial "point" represented by the value of P is actually a square in $\mathbb{R}^2$. As k increases, the sizes of the squares decrease. In fact, as k tends to infinity, the value of P will tend to represent a specific point in $\mathbb{R}^2$.

The space-filling curve created by the values of the Peano key, P, is in the shape of a recursive N. Figure 2 illustrates the N-curve, using a grid of 1024 points. This figure displays the self-similarity feature of fractal images.

The N-curve passes once and only once through each point in space, points being defined as squares whose size is determined by the number of digits carried in the latitude and longitude coordinates. The order of points on the curve (Peano order) is largely preserving of geographic contiguity. Thus, Peano order facilitates proximity searches. Peano order involves a few geographic discontinuities, such as the jump from point 516 to point 517 in figure 2, as does any mapping from $\mathbb{R}^2$ to $\mathbb{R}^1$.

In the specific application we envision here, economic establishments are arranged on a list in Peano order by means of their latitude and longitude coordinates. Probability samples of the establishments may be drawn systematically from the ordered list. Because the earth's coordinate system is stable, there is no ambiguity in determining the list position of new establishments. Thus, they may be subjected to sampling too.

To illustrate this application, see figure 3 which displays a chain of retail establishments in the United States. Each establishment is described by a double-letter code. This code in natural lexicographic order signifies the Peano order of the establishments.

In the next section, we describe a sample maintenance system that is based upon the establishments' Peano order.

Figure 2. Peano Order Based on 1024 Points

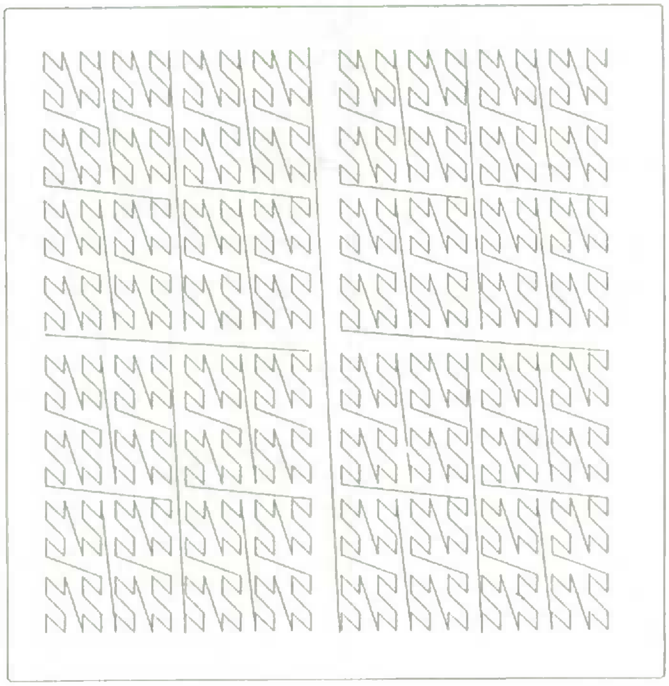
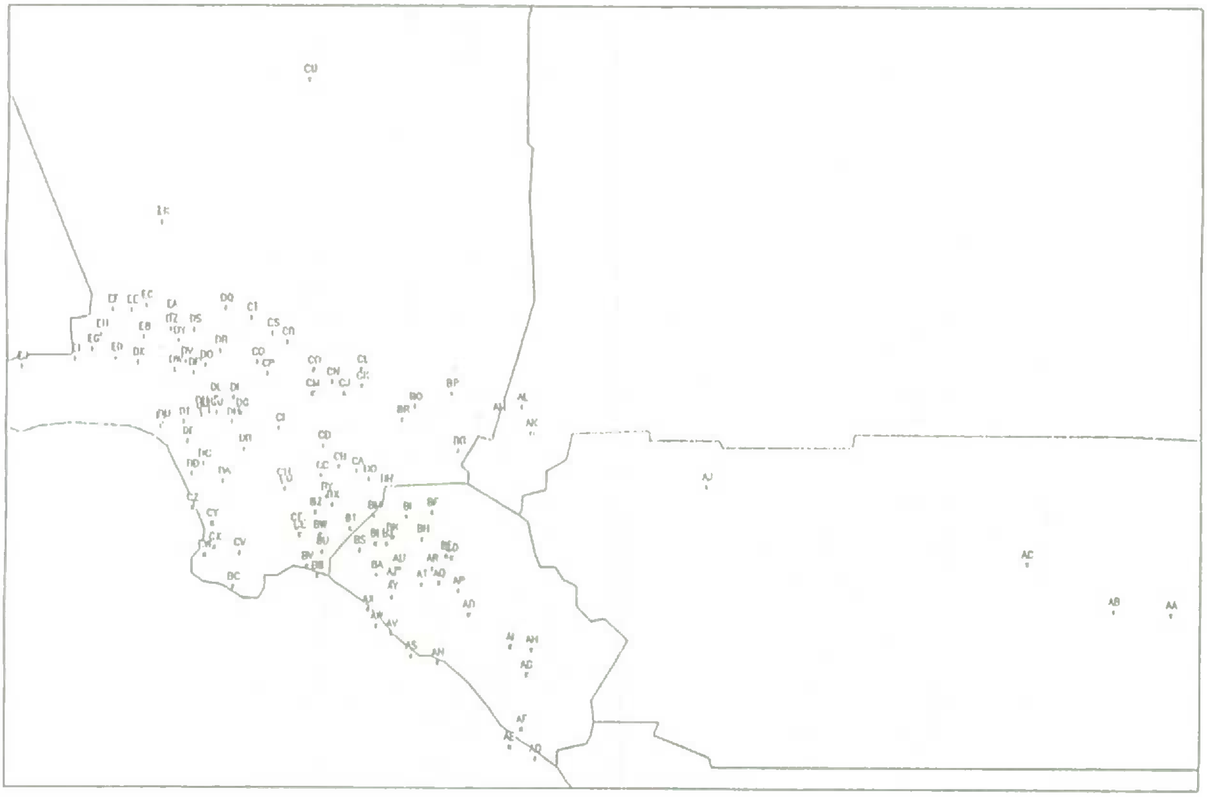


Figure 3. Chain of Retail Establishments in Peano Order



4. RULES FOR MAINTAINING THE SAMPLE

We describe a system for maintaining samples of retail stores, taking proper account of births, deaths, scanning conversions, and other changes in the status of the retail store universe. As stated earlier, we developed the system for applications at the A. C. Nielsen Company.

We consider a given and arbitrary sampling stratum, say of size N, and assume the universe of stores in the stratum is arranged in Peano order. For example, a stratum might include all stores in a given metropolitan market, such as Vancouver or Montreal. Ordering by Peano key values will turn out to be especially well-suited to the maintenance system that follows. Other ordering schemes may be considered for this work so long as they are stable across time and effectively map $\mathbb{R}^2$ to $\mathbb{R}^1$ in such fashion as to preserve geographic contiguity and to assign all birth stores a unique position in the ordering.

We assume an original sample is selected systematically with equal probability from the ordered list of stores at time $t = 0$. Let U_{ij} denote the j-th store in the i-th possible systematic sample, for $i = 1, \dots, k$ and $j = 1, \dots, n_i$, where k is the sampling interval and n_i is the size of the i-th possible sample. If $N = nk + r$, $r < k$, then r samples will be of size $n_i = n + 1$ and $k - r$ samples of size n. In what follows, we shall also use the subscript "i" to represent the sample actually selected.

Let P_{ij} denote the Peano key value associated with U_{ij}. Let P_L and P_U denote the smallest and largest possible Peano key values within the market under study. Thus,

$$P_L \le P_{11} < P_{21} < \dots < P_{k1} < P_{12} < \dots < P_{ij} < \dots < P_{kn_k} \le P_U.$$

Note that we are assuming each store possesses a unique geographic location and thus a unique Peano key value.

Let Y_{tij} denote the value of some characteristic of U_{ij} at time t. A standard, unbiased estimator of the population total, Y_t, is

$$\hat{Y}_{ti} = k \sum_{j=1}^{n_i} y_{tij} ,$$

while the ratio estimator is given by

$$\hat{Y}_{Rti} = \hat{Y}_{ti} X_t / \hat{X}_{ti} ,$$

where the X-variable is a measure of size and X_t and $\hat{X}_{ti}$ are analogous to Y_t and $\hat{Y}_{ti}$, respectively.

Define N Peano key segments, S_{ij}, by partitioning the range $[P_L, P_U]$ at the N store values P_{ij}. We let $S_{ij} = [P_{ij}, P_{i+1,j})$, where it will be understood that $P_{k+1,j}$ represents $P_{1,j+1}$. A special definition is needed for the final segment. We define $S = [P_{kn_k}, P_U] \cup [P_L, P_{11})$ so that the entire Peano range $[P_L, P_N]$ is covered by the N segments. This special definition, which treats the Peano range as if it were on a circle, is needed later to guarantee that all store births are given a nonzero probability of selection. Alternative segmentation schemes may be used without defeating the statistical properties of the maintenance system.

Our maintenance scheme is based upon the Peano key segments. The basic idea is to view the systematic selection process as applying to the segments, with subsampling of stores within the selected segments. Thus, as a formal matter, the segment is the primary sampling unit (PSU), not the store. Of course, as of the time of initial sample selection, there is, by construction, only one store per segment.

4.1 Birth Sampling

At a future point in time, say t', one or more new stores may open for business. Each new store will be assigned its unique Peano key value, and this value will be an element of one and only one Peano key segment. The Peano key permits us to automatically place new stores in their correct and unique positions on the ordered universe list.

The simplest possible rule for sampling births is the following:

> Rule 1. A birth store is selected into the sample if and only if its Peano key value is an element of a selected Peano key segment. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

Given this rule, a birth store is selected with probability 1/k. This occurs because its segment, which is unique, is selected with probability 1/k. Unfortunately, Rule 1 does not provide good control of the sample size over time.

To control the sample size, we advocate some form of subsampling within PSU's. Let U_{ij1}, U_{ij2}, ..., $U_{ijB_{ij}}$ denote the stores in segment S_{ij}. The original store is now labeled U_{ij1}, whereas U_{ij2}, U_{ij3}, ..., $U_{ijB_{ij}}$ are the birth stores in Peano order. The number, $B_{ij} - 1$, of births in any given segment will be 0, 1, or 2 in most applications. Then, we may subsample as described in the following alternative rule.

> Rule 1A. A birth store will be subjected to subsampling if and only if its Peano key value is an element of a selected Peano key segment. Associate with U_{ij1}, U_{ij2}, ..., $U_{ijB_{ij}}$ the probabilities p_{ij1}, p_{ij2}, ..., $p_{ijB_{ij}}$, where $p_{ijb} > 0$ and $\Sigma\, p_{ijb} = 1$. Now choose one of the stores according to this probability measure. Subsampling is independent from one selected segment to the next. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

The probabilities in Rule 1A may be equal or unequal. If unequal, they may be defined in proportion to some preliminary measure of size, or defined so as to accelerate or retard the replacement of the sample.

We observe that our principal maintenance objectives are well-satisfied by Rule 1A. First, the rule maintains geographic balance over time because there is always one unit selected from each of the originally selected segments, which themselves were geographically balanced by virtue of the systematic sampling design. Second, the rule maintains a constant sample size over time because there is always one and only one store selected from each of the originally selected segments. Third, the rule is in accord with strict principles of probability sampling, whereby probabilities of inclusion are known and nonzero, and thus unbiased estimators of population totals are available. Finally, by appropriate choice of the p_{ijb}, we may control distortion in year-to-year trends.

The unconditional probabilities of selection are given by

$$\pi_{ijb} = k^{-1}\, p_{ijb}$$

for $b = 1, \ldots, B_{ij}$. That is, π_{ijb} is equal to the probability of selecting the PSU times the conditional probability of selecting the store, given the selected PSU.

Let $Y_{t'ijb}$ denote the value of the unit U_{ijb}, and let $Y_{t'ij+}$ denote the total for the (i,j)-th PSU. Then, the unbiased estimator of the population total $Y_{t'}$ is given by

$$\hat{Y}_{t'i} = \sum_{j=1}^{n_i} y_{t'ijb} / \pi_{ijb} \quad ,$$

where $y_{t'ijb}$ is the value of the single unit selected from the (i,j)-th selected segment, with variance

$$\mathrm{Var}\{\hat{Y}_{t'i}\} = \frac{1}{k} \sum_{i=1}^{k} (k \sum_{j=1}^{n_i} Y_{t'ij+} - Y_{t'})^2 + k \sum_{i=1}^{k} \sum_{j=1}^{n_i} \sigma^2_{t'ij} \quad , \qquad (1)$$

where

$$\sigma^2_{t'ij} = \sum_{b=1}^{B_{ij}} p_{ijb} \left(\frac{Y_{t'ijb}}{p_{ijb}} - Y_{t'ij+}\right)^2 \quad .$$

The first term on the right side of (1) is the variance due to the sampling of segments. This is the original variance in the sense that it is the variance expression that applied at the time of original sample selection. The second term on the right side is the variance due to subsampling within segments. Note that $\sigma^2_{t'ij}$ vanishes for any segment in which birth subsampling has not occurred. Note also that the subsampling scheme achieves its minimum variance when, for each given i and j, the probabilities p_{ijb} are defined to be proportional to $Y_{t'ijb}$. In this case, the within component of variance vanishes. For any real application, however, this proportionality condition will be satisfied only approximately.

As usual, a first-order Taylor series approximation may be used to discover the variance of the ratio estimator. See Wolter (1986) for appropriate techniques to estimate the variance of both the unbiased estimator, $\hat{Y}_{t'i}$, and the ratio estimator, $\hat{Y}_{Rt'i}$.

As time passes, it will be necessary to periodically update the sample to reflect additional births and other changes in the universe. It may be desirable to schedule the updating at regular intervals of time, so as to facilitate management of the work. I will refer to these intervals as update cycles. Such cycles may occur monthly, bimonthly, quarterly, or at whatever interval makes sense in a particular application. Factors to consider in establishing the frequency of the updating cycles include cost of the updating process; desired accuracy of the estimators of level and trend; and perceptions of the customers or users of the data.

Generally speaking, more frequent updating will cost more, achieve greater accuracy, and be perceived better by customers than less frequent updating.

For an update cycle at any future time t', Rules 1 or 1A may be used to maintain the sample. New stores are always placed automatically in their correct segment, by their Peano key values, and the subscript b reflects this order at each cycle. To explicitly reflect these ideas, we should have further subscripted the U's, B's, p's, and π's by time, but we avoided doing so as a notational convenience. The expressions for the estimators of total, $\hat{Y}_{t'i}$ and $\hat{Y}_{Rt'i}$, and their variances remain valid for each t'.

4.2 Updating for Deaths

Rules for maintaining a sample over time must obey an important general principle. They must treat equally both selected and nonselected units. In the case of deaths, this principle implies that all deaths, both those in and out of the sample, must be handled in the same fashion in any sample updating process. If this principle is not followed, the resulting estimators will be biased, and the bias may accumulate over time.

In what follows, we describe procedures for death updating that follow this essential principle. There are two cases to consider: (i) deaths are not known on a universe basis, (ii) deaths are known on a universe basis.

For case (i), we suggest Rule 2.

> **Rule 2. All deaths in the sample will be known. They should remain in the sample but be set to zero (i.e., y = 0) at the time of an update cycle.**

This rule permits unbiased estimation of the universe population totals. Deaths cause the estimator variances to increase, and estimators of variance will properly reflect this increase, provided the deaths are retained in the sample with zero values.

For case (ii), we suggest Rule 3.

> **Rule 3. Remove all deaths from the universe at the time of the next update cycle. Subject only the remaining live cases to sampling, including births.**

Rule 3 will cause the store count B_{ij} to change in segments where deaths have occurred, unless births exactly offset deaths. In fact, the B's and p's will necessarily change in segments where there are deaths and no births. As a consequence, a replacement store will necessarily be selected within a given segment whenever the sample store from the segment has died, and a replacement store may be selected even when the sample store is alive and well.

Two additional issues must be addressed in handling deaths. The first issue concerns the coordination of birth and death updating. Store births and deaths will occur naturally at irregular intervals, depending upon business conditions and population growth. In some time periods, neither births nor deaths will occur. In other time periods, births may occur but not deaths, or vice versa. While in other periods, both deaths and births will occur. In theory, it would be possible to employ different update cycles for grocery store births and deaths. For example, one might update bimonthly for both births and deaths, but in alternating months. This approach may have advantage in leveling the work load over time. On the other hand, alternating cycles may tend to defeat the ability of the sample to properly measure trends, creating a sawtooth pattern in the store time series as first births are introduced, then deaths dropped, then births, deaths, and so on. On balance, we recommend coincident sample updating for births and deaths so as to preserve trends.

The second issue concerns the handling of deaths during the period from their actual occurrence until the next update cycle. This issue arises only if the frequency of the updating process is less than that of the data-collection process. If the two processes are coincident, then there are no new problems. If updating is the less frequent, then there are two alternatives:

a) drop the deaths from the sample as soon as they are known to us (to be more precise statistically, this means the deaths are included in the sample with a value of zero)

b) continue the deaths in the sample by imputing for them until the time of the next update cycle.

Alternative a) is the simplest, cleanest way of proceeding. Aside from the problem of births, it is unbiased and permits correct variance estimators. Because of the birth problem, however, this alternative may have a negative effect on the ability of the sample to properly measure trends. As deaths occur during the first weeks of an update cycle, one can imagine a slight decline in the store time series, not because of fundamental change in economic conditions, but simply because the sample reflects deaths and not births. Alternative b) provides a short-term fix to the problem of properly measuring trends. The essential notion here is that by imputing for deaths, we implicitly make a correction for any births that have occurred since the last update cycle. This fix is not particularly elegant, and it is difficult to frame a rigorous, unassailable technical justification for it. On the other hand, history has shown that populations of economic establishments tend to be stable in the short run. Deaths are often associated with or are compensated by births, with the net size of the population remaining approximately level in the short run. The United States Bureau of the Census has used this alternative in its wholesale trade survey, with quarterly update cycles and monthly data collection. See Wolter et al (1976).

4.3 Scanning Conversions

In this final subsection, we present sample maintenance rules for handling stores that convert from nonscanning to scanning, and vice versa. Of course, this particular type of universe dynamic does not arise in surveys that utilize other data collection technologies.

First, we treat conversions to scanning. There are two principal cases to consider: (i) scanning status is known for all stores prior to sampling; (ii) scanning status is not known prior to sampling, but is observed after sampling for the selected stores only.

Case (i) is relatively easy to handle. Here is a natural rule:

> Rule 4. Do not subject nonscanning stores to sampling. Sample only from the subuniverse of scanning stores. As a given nonscanning store converts to scanning, then treat it as a birth, subjecting it to birth sampling. Prior to conversion, nonscanning stores shall be represented in the universe by utilizing imputation or other missing data techniques.

Given this rule and the prior data (i.e., scanning status) it assumes, the entire survey budget may be allocated to the sample of scanning stores. None of the sample resources need be committed to nonscanning stores. Unfortunately, this desirable property does not hold for case (ii).

To address case (ii), some additional notation is needed. Let A denote the set of scanning stores and B the set of nonscanning stores, where $A \cup B$ spans the entire universe. Set s denote the selected sample of stores, and let $s_A = s \cap A$ and $s_B = s \cap B$.

By assumption, s_A and s_B are not observed until after initial field work is completed. Obviously, all of these sets vary with time, but we suppress explicit time subscripts to simplify the notation.

Sample s_A should be maintained by rules presented elsewhere in this paper for births and deaths. New rules are required to handle s_B. Here is an illustrative rule that treats the stores in s_B as nonrespondents.

> Rule 5. At time t, impute for store $U_{ijb} \in s_B$ the value $\hat{y}_{tijb} = x_{tijb}\, y_{At} / x_{At}$, where x_{tijb} is the value of an auxiliary variable for store U_{ijb}, y_{At} is the sample s_A total for the estimation variable, and x_{At} is the corresponding total for the auxiliary variable. Alternatively,imputation may occur by means of substitution, hot deck/matching, or other means. Now, act as if the data set is complete, applying standard estimators of the survey parameters of interest. At the time U_{ijb} converts to scanning, it shall be deleted from s_B and joined to s_A, and the estimation shall still be performed by means of the standard estimators applied to the completed data set.

Given Rule 5, the effective sample size is reduced because of imputation variance associated with the $\hat{y}_{tijb}$. Substitution maintains a larger effective sample size than the other rules, but is clearly the most expensive to implement. All rules require limited field work on a continuous basis to monitor the scanning status of $U_{ijb} \in s_B$.

As an alternative to missing data techniques, we may observe the nonscanning stores using an alternative mode of data collection. Depending upon the data to be collected, this could involve a store audit or an interview conducted with store personnel by telephone, mail, or in person. This alternative would likely be more accurate than the imputation-based methods, yet additional cost and time may be involved, as well as burden associated with the management and control of two data collection methodologies.

Finally, we treat conversions of sample stores from scanning to nonscanning. Such conversions are likely to be relatively small in number and are treated here only for completeness.

Let $U_{ijb} \in s_A$, i.e., i is a scanning store in the sample. Note that U_{ijb} may be either a store that has scanned since being selected into the sample, or a store that converted to scanning after originally entering the sample as a nonscanner under Rule 5.

> Rule 6. At the time U_{ijb} converts to nonscanning, it shall be deleted from s_A, joined to s_B, and subsequently handled by missing data techniques, as in Rule 5. Standard formulae shall be applied to the completed data set. To simplify processing and field work, the method selected shall be identical to the method selected to handle conversions from nonscanning to scanning.

In the bizarre instance in which a store flip-flops repeatedly between scanning and nonscanning, one may handle the store by sequentially applying Rule 5 or 6, as the case may be, each time updating the sets s_A and s_B.

REFERENCES

Ernst, L. (1989) "Weighting Issues for Longitudinal Household and Family Estimates," in Panel Surveys, edited by Kaspryzk, D., Duncan, G., Kalton, G., and Singh, M. P., Wiley, NY.

Hanson, R. H. (1978) "The Current Population Survey: Design and Methodology," Technical Paper 40, United States Bureau of the Census, Washington, DC.

Laurini, R. (1987) "Manipulation of Spatial Objects by a Peano Tuple Algebra," University of Maryland Technical Report CS-TR-1893, College Park, MD.

Peano, G. (1908) "La Curva di Peano nel 'Formulario Mathematico.'" In "Opere Scelte di G. Peano," pp. 115-116, Vol. I. Edizioni Cremonesi, Roma, 1957.

Progressive Grocer (1989) "56th Annual Report of the Grocery Industry 1989," Vol. 68, No. 4, Part 2, Stamford, CT.

Rao, J. N. K. and Graham, J. R. (1964) "Rotation Designs for Sampling on Repeated Occasions," Journal of the American Statistical Association, 59, 492-509.

Saalfeld, A.; Fifield, S.; Broome, F.; and Meixler, D. (1988) "Area Sampling Strategies and Payoffs using Modern Geographic Information System Technology," unpublished paper, United States Bureau of the Census, Washington, DC.

Sirken, M. (1970) "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65, 257-266.

Wolter, K. M. (1979) "Composite Estimation in Finite Population," Journal of the American Statistical Association, 74, 604-613.

Wolter, K. M. (1986) Introduction to Variance Estimation, Springer Verlag, NY.

Wolter, K. M. et al (1976) "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys, "Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

MARGINAL AND APPROXIMATE CONDITIONAL LIKELIHOODS FOR SAMPLING ON SUCCESSIVE OCCASIONS

D.R. Bellhouse[1]

SUMMARY

Marginal and approximate conditional likelihoods are given for the correlation parameters in a normal linear regression model with correlated errors, both under a fixed regression parameter assumption and under a random coefficients regression model. These likelihoods may be evaluated using state space models. This general likelihood approach is applied to obtain marginal and conditional likelihoods for the correlation parameters in sampling on successive occasions under both simple random sampling on each occasion and more complex surveys.

KEY WORDS: Likelihood inference, Sampling in time, ARMA models, State space models.

1. INTRODUCTION

Marginal likelihoods were introduced as a general method for eliminating nuisance parameters from the likelihood function (Fraser, 1967; Kalbfleisch and Sprott, 1970). Cox and Reid (1987) introduced approximate conditional likelihoods which also address this problem. They argued that the approximate conditional likelihood was preferable to the profile likelihood obtained by replacing the nuisance parameters in the likelihood by their maximum likelihood estimates when the parameters of interest are given. Bellhouse (1990) established the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. Following on the work of Cox and Reid, Cruddas et al. (1989) obtained an approximate conditional likelihood for the correlation parameter in several short series of autoregressive processes of order one with common variance and autocorrelation parameters. Based on a simulation study, Cruddas et al. (1989) showed that the estimate based on the approximate conditional likelihood has a much smaller bias and better coverage properties of the confidence interval than the maximum likelihood estimate from the profile likelihood.

A situation similar to the one studied by Cruddas et al. (1989) appears in sampling on successive occasions in sample surveys. In order to reduce the response burden, individuals in a survey are retained in the sample for relatively short periods of time. For any occasion on which the survey is carried out, the sample consists of some individuals who have been previously surveyed on some past occasion or occasions, and some who are new to the survey for the first time. The sample measurements on an individual are usually modelled by an autoregressive moving average process (ARMA); see Binder and Hidiroglou (1988) for a review of the application of time series models to sampling on successive occasions. Moreover, because of the response burden, the observed time series for an individual is short. If the model means on each occasion are assumed to be different, then the dimension of the parameter space increases with time so that the maximum likelihood estimates of the parameters can be biased and inconsistent. Consequently, it is of interest to obtain marginal and approximate conditional likelihoods under ARMA models.

The marginal and approximate conditional likelihoods for the correlation parameters in a normal model are obtained in section 2. The general results of section 2 are illustrated in section 3 by applying the results to sampling on successive occasions assuming simple random sampling. In section 4, several methods are given to apply these likelihood methods to complex surveys.

2. MARGINAL AND APPROXIMATE CONDITIONAL LIKELIHOODS FOR CORRELATION PARAMETERS UNDER A NORMAL MODEL

For the linear model

$$y = X\beta + \varepsilon \tag{1}$$

with error vector $\varepsilon \sim N(0,\sigma^2\Omega)$, where Ω is the correlation matrix, the log-likelihood for β, σ^2 and Ω is given by

$$l(\beta,\sigma^2,\Omega) = -\{m \ln\sigma + (\ln|\Omega|)/2 + (y-X\beta)^T\Omega^{-1}(y-X\beta)/2\sigma^2)\} \tag{2}$$

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B9.

The vector of observations y is of dimension $m \times 1$ and the vector of regression coefficients β is $p \times 1$ so that X is $m \times p$. For a given value of Ω,

$$\hat{\beta} = (X^T\Omega^{-1}X)X^T\Omega^{-1}y$$

and

$$s^2 = y^T\Omega^{-1}y - y^T\Omega^{-1}X\ (X^T\Omega^{-1}X)^{-1}\ X^T\Omega^{-1}y \tag{3}$$

are jointly sufficient for β and σ^2.

A marginal likelihood for Ω is obtained by making a transformation of the data y to the sufficient statistics $\hat{\beta}$ and s^2 and the ancillary statistic

$$a = \Omega^{-1/2}(y - X\ (X^T\Omega^{-1}X)^{-1}\ X^T\Omega^{-1}y)/s,$$

where $\Omega^{-1/2}$ is the $m \times m$ dimensional matrix such that $\Omega^{-1} = \Omega^{-1/2}\Omega^{-1/2}$. The marginal likelihood of Ω is the marginal distribution of the ancillary a times the product of the differentials da_i, $i=1, \ldots, m$. See Kalbfleisch and Sprott (1970, eqs. 6 and 10) for a general discussion and a general expression for Πda_i. Bellhouse (1978) and, later independently Tunnicliffe Wilson (1989), showed that the marginal likelihood for Ω under the normal model is given by

$$l_M(\Omega) = \{\ |\Omega|^{1/2}\ |X^T\Omega^{-1}X|^{1/2} s^{m-p}\ \}^{-1}. \tag{4}$$

Note that (3) is proportional to the maximum likelihood estimate of σ^2 given Ω and that $s^2(X^T\Omega^{-1}X)^{-1}$ is proportional to the estimated variance-covariance matrix of the maximum likelihood estimate of β given Ω. Then (4) can be written as

$$L_M(\Omega) = \frac{|\text{est var}(\hat{\beta})|^{1/2}}{s^m|\Omega|^{1/2}}. \tag{5}$$

To obtain an approximate conditional likelihood, it is first necessary to transform the parameters to achieve parameter orthogonality between the parameters of interest and the nuisance parameters, which now may depend on the parameters of interest. Sets of parameters are orthogonal if the associated information matrix is block diagonal, with each block as the information matrix for each parameter set. The conditional likelihood is related to the distribution of the data y conditional on the maximum likelihood estimate of the nuisance parameters for fixed values of the parameters of interest. The approximate conditional likelihood is obtained by applying two approximations to this conditional distribution. See Cox and Reid (1987, section 4.1) for a discussion of the derivation. For example, let Θ be the vector of parameters of interest and let Λ, possibly depending on Θ, be the vector of nuisance parameters orthogonal to Θ. The full likelihood of the data for parameters Θ and Λ is denoted by $L(\Theta,\Lambda)$ and the profile likelihood for Θ, $L(\Theta,\hat{\Lambda})$ is the likelihood with Λ replaced by its maximum likelihood estimate. The approximate conditional likelihood for Θ is

$$L(\Theta,\hat{\Lambda})\ |\ I(\Theta,\hat{\Lambda})\ |^{1/2},$$

where $I(\Theta,\hat{\Lambda})$ is the observed information matrix for Λ at a fixed value of Θ. See Cox and Reid (1987, eq. 10).

Following Cruddas et al. (1989), Bellhouse (1990) suggested, for model (1), the parameter transformation $\lambda = \ln\sigma + (\ln|\Omega|)/(2m)$ leaving β the same. The log-likelihood under the new parameterization is denoted by $l(\beta,\lambda,\Omega)$ and can be obtained from (2). If the entries of Ω are functions of a parameter Φ, then the nuisance parameters λ and β are each orthogonal to Ω, i.e.

$$-\frac{1}{m}\ E[\frac{\partial^2 l(\beta,\lambda,\Omega)}{\partial\Phi\partial\lambda}] = 0$$

and

$$-\frac{1}{m}\ E[\frac{\partial^2 l(\beta,\lambda,\Omega)}{\partial\Phi\partial\beta}] = 0,$$

when each entry of Ω is a continuous and differentiable function of Φ. Moreover, in this case the approximate conditional likelihood for Ω, $L_C(\Omega)$ is the same as the marginal likelihod $L_M(\Omega)$, given by (4) or (5). See Bellhouse (1990) for details.

The marginal and approximate conditional likelihod in (4) or (5) can be evaluated at any Ω using state space models in the approach of Harvey and Phillips (1979). For any given Ω, once the recursions to estimate β and σ^2 are complete, the value of s^2 and $|\Omega|^{1/2}$ can be calculated from Harvey and Phillips (1979), eqs. 5.6 and 6.6, and 4.3 respectively). It is then necessary only to obtain $X^T\Omega^{-1}X$ and its determinant. The value of $X^T\Omega^{-1}X$ may be obtained from the final step in the recursive equations of Harvey and Phillips (1979, eq. 3.4).

Suppose in model (1) that β is a random vector modelled by $\beta = W\delta + u$, where W is a $p \times q$ matrix of known values, δ is a $q \times 1$ vector of parameters, and $u \sim N(0, \gamma^2\Gamma)$, independent of ε. Under the composite model $y = XW\delta + Xu + \varepsilon$, the log-likelihood for $\delta, \Omega, \Gamma, \gamma^2$, and $\kappa = \sigma^2/\gamma^2$, denoted by $l(\delta, \kappa, \gamma^2, \Gamma, \Omega)$, is given by (2), with Ω replaced by $\kappa\Omega + X\Gamma X^T$ and $X\beta$ replaced by $XW\delta$. Likewise, the marginal likelihood, denoted by $L_M(\kappa, \Gamma, \Omega)$, is given by (4) and (3), with X replaced by XW and Ω replaced by $\kappa\Omega + X\Gamma X^T$. This yields

$$l_M(\kappa,\Gamma,\Omega) = \{|\kappa\Omega + X\Gamma X^T|^{1/2}|(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW|^{1/2} g^{m-q}\}^{-1}, \tag{6}$$

where

$$g = y^T(\kappa\Omega + X\Gamma X^T)^{-1}y$$
$$- y^T(\kappa\Omega + X\Gamma X^T)^{-1}XW((XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW)^{-1}(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}y.$$

Now the dimension of Ω may be large in comparison to Γ; this can be the case in sampling on successive occasions. As an alternate approach, one could take the likelihood implied by (2), multiply it by the distribution for β, and integrate over β to obtain the likelihood for the parameters under the random model. This will yield matrices of the same dimension as Γ.

3. SIMPLE RANDOM SAMPLING ON SUCCESSIVE OCCASIONS

3.1 Rotation Sampling

Consider a finite population of N units which has been sampled on k occasions by one-level rotation sampling. Let y_{tj} denote the measurement on the j^{th} population unit taken on the t^{th} occasion, $j=1, \dots, N$ and $t=1, \dots, k$. To begin with, it is assumed that any two units, say j and j' are independent, but that the same unit across time is correlated. In particular, assume that for any j,

$$(y_{1j}, y_{2j}, \dots, y_{kj})^T \sim N(\mu, \sigma^2\Omega_k), \tag{7}$$

where Ω_k is a $k \times k$ correlation matrix and where μ is the $1 \times k$ vector of fixed means $(\mu_1, \mu_2, \dots, \mu_k)^T$.

The notation of Bellhouse (1989) is used to describe the sampling scheme. On any occasion, c rotation groups are sampled. Rotation group r, denoted by G_r, consists of m_r sample units, $r = 1, 2, \dots, k + c - 1$. On occasion t, the sample consists of the units in $G_t, G_{t+1}, \dots, G_{t+c-1}$, so that the total sample size on occasion t, $n_t = m_t + m_{t+1} + \dots + m_{t+c-1}$. Each rotation group is chosen by simple random sampling without replacement from previously unchosen units in the population. The total sample size over all k occasions is $m = n_1 + n_2 + \dots + n_k$.

Suppose G_r first appears in the sample on occasion u and last appears on occasion v; u is either 1 or r and v is either $r + c - 1$ or k. The total number of occasions on which a unit in G_r is present in the sample is $b = v + 1 - u$. Let $\bar{y}_{u,r}, \dots, \bar{y}_{v,r}$ be the sample means or elementary estimates for G_r on occasions $u, u + 1, \dots, v - 1, v$ respectively. Then under model (7), the contribution of G_r to the log likelihood in (2) is

$$- \{bn_r \ln\sigma + (n_r/2)\ln(|\Omega_r|) +$$
$$[n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1)\,\mathrm{tr}(\Omega_r^{-1}S_r)]/(2\sigma\}, \tag{8}$$

where x_r^T is the 1 x b vector $(\bar{y}_{u,r} - \mu_u, \bar{y}_{u+1,r} - \mu_{u+1}, \ldots, \bar{y}_{v-1,r} - \mu_{v-1}, \bar{y}_{v,r} - \mu_v)$, where S_r is the b x b matrix of sums of squares and cross products of observations within the rotation group, and where Ω_r is the b x b correlation matrix on the observations on a single unit within the rotation group. By the independence assumption, the full log likelihood is obtained by summing (8) over all rotation groups.

Given Ω, or equivalently $\Omega_1, \ldots, \Omega_{k+c-1}$, expressions for the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$, for μ and σ^2 respectively, may be found. Likewise, $V(\hat{\mu})$, the estimated variance-covariance matrix of $\hat{\mu}$ may be obtained. This is illustrated for a first-order autoregressive process in section 3.2. Then the marginal likelihood for the parameters in $\Omega_1, \ldots, \Omega_{k+c-1}$ is given by (4) with the expressions in (4) given by

$$|\Omega|^{1/2} = \prod_{r=1}^{k+c-1} \Omega_r,$$

$$|X^T\Omega^{-1}X|^{1/2} = V(\hat{\mu})/s^k,$$

$$s^2 = \sum_{r=1}^{k+c-1} \{(n_r\hat{x}_r^T\Omega_r^{-1}\hat{x}_r + (n_r - 1)\ tr(\Omega_r^{-1}S_r)\}, \tag{9}$$

and $p = k$, where $\hat{x}_r$ is x_r with the μ's in x_r replaced by their maximum likelihood estimates.

3.2 First-Order Autogressive Processes

Consider an autoregressive model which allows independence between different units but correlation within a unit over time. In particular, assume the first-order autoregressive model

$$y_{tj} = \mu_t + \Phi\ (y_{t-1,j} - \mu_{t-1}) + \varepsilon_{tj}, \tag{10}$$

where $\varepsilon_{ij} \sim N(0,\sigma^2)$ for $t = 1, \ldots, k$ and $j = 1, \ldots, N$, and where the ε's are mutually independent. Model (9), essentially Patterson's (1950) model, is a special case of (7). As in section 3.1, the vector of regression parameters $\beta = (\mu_1, \ldots, \mu_k)^T$. When the data vector **y** contains the measurements on each unit grouped by all the occasions on which it was sampled as in the rotation sampling description of section 3.1, the correlation matrix Ω, now a function of Φ, can be written as a direct sum of matrices, each of which are the correlation matrices of a first-order autoregressive process.

The following notation, similar to Patterson (1950), is used to denote various sample sizes, means and sums of squares and cross products (corrected for the appropriate mean) for occasion t:

π_t = the proportion of units on occasion t that are matched with units from the previous occasion (t-1);

n_t = the number of units sampled on occasion t;

$\bar{y}_t'$ = the mean of the units on occasion t that are matched with units from the previous occasion (t-1);

$\bar{y}_t''$ = the mean of the units on occasion t that are unmatched with units from the previous occasion (t-1);

$\bar{y}_t$ = the mean of all the units on occasion t;

$\bar{x}_t'$ = the mean of the units on occasion t that are matched with units from the following occasion (t+1);

syy_t' = the sum of squares among units on occasion t which are matched with units from the previous occasion (t-1);

syy_t'' = the sum of squares among units on occasion t which are unmatched with units from the previous occasion (t-1);

sxx_t' = the sum of squares among units on occasion t which are matched with units from the following occasion (t+1);

sxy_t' = The sum of squares among all the units on occasion t;

syy_t = the sum of cross products for measurements on sample units from occasion t matched with sample units from $t-1$.

Under the special case of model (10), and after much algebra, it may be shown that (8) summed over all rotation groups r, the log-likelihood for the data reduces to

$$l(\mu_1, \ldots, \mu_k, \sigma^2, \Phi) = -m \ln\sigma + (d/2)\ln(1-\Phi^2) - \{A(\mu,\Phi) + B(\Phi)\}/(2\sigma^2), \tag{11}$$

where d is the distinct number of units sampled (irrespective of the number of occasions on which a unit is sampled) and m is the total sample size $(n_1 + \ldots + n_k)$. Further in (11),

$$A(\mu,\Phi) = (1-\Phi^2)n_1(\bar{y}_1 - \mu_1)^2$$

$$+ \sum_{t=2}^{k} [\pi_t n_t\{\bar{y}_t' - \mu_t - \Phi(\bar{x}_{t-1}' - \mu_{t-1})\}^2 + (1 - \pi_\tau)n_t(1 - \Phi^2)(\bar{y}_t'' - \mu_t)^2] \tag{12}$$

and

$$B(\Phi) = (1-\Phi^2)\ syy_1 + \sum_{t=2}^{k} \{\Phi^2\ sxx_{t-1}' - 2\Phi\ sxy_t' + syy_t' + (1-\Phi^2)\ syy_t''\}. \tag{13}$$

For any given value of Φ the maximum likelihood estimator is $\hat{\mu} = G^{-1}z$ and $\hat{\sigma}^2 = \{A(\hat{\mu}, \Phi) + B(\Phi)\}/m$, where $A(\hat{\mu},\Phi)$ is (12) with μ replaced with its maximum likelihood estimate and where G is a symmetric $k \times k$ band matrix of bandwidth 3 and z is a $k \times 1$ vector. The nonzero entries of G are

$$g_{tt} = \pi_t n_t + (1 - \pi_t)n_t(1 - \Phi^2) + \pi_{t+1}n_{t+1}\Phi^2, \text{ for } t = 1, \ldots, k \tag{14}$$

and

$$g_{t,t+1} = -\pi_{t+1}n_{t+1}\Phi, \text{ for } t = 1, \ldots, k-1, \tag{15}$$

where $\pi_1 = \pi_{k+1} = 0$. The entries of z are

$$z_t = \pi_t n_t(\bar{y}_t' - \Phi\bar{x}_{t-1}') + (1 - \pi_t)n_t\bar{y}_t''(1 - \Phi^2) - \pi_{t+1}n_{t+1}(\bar{y}_{t+1}' - \Phi\bar{x}_t'), \tag{16}$$

for $t = 1, \ldots, k$, where $\pi_1 = \pi_{k+1} = 0$ and $\bar{y}_1'' = \bar{y}_1$. The vector of estimated means $\hat{\mu}$ is unbiased for μ under model (10) and its variance-covariance matrix is $\sigma^2 G^{-1}$. It follows from (4) or (5) that the marginal and approximate conditional likelihood for Φ is

$$L_M(\phi) = \frac{(1 - \delta^2)^{d/2}}{\{A(\hat{\mu},\phi) + B(\phi)\}^{(m-k)/2}|G|^{1/2}}. \tag{17}$$

3.3 Random Model Means

The discussion in section 3.1 and 3.2 has ignored the possibility of a relationship between the means for each occasion. The means for each occasion are the ultimate quantity of interest, and much information may be lost if the relationship between the means over time is ignored. Blight and Scott (1973), for example, note that the survey means on successive occasions are often correlated and assume, in addition to (10), that

$$\mu_t - \xi = \phi(\mu_t - \xi) + u_t, \tag{18}$$

where $u_t \sim N(0,\gamma^2)$ and where the u's are mutually independent. Model asumptions such as (18) can be incorporated in the estimation procedures in at least two ways.

The first method is to use the full likelihood approach. Under the model defined by (10) and (18), the log-likelihood for the data becomes

$$L(\xi,\gamma^2,\Theta,\phi,\kappa) = -m \ln\gamma + (k/2) \ln\kappa + (d/2) \ln(1 - \Phi^2) + (1/2) \ln(1 - \phi^2)$$
$$- \{A(\hat{\mu},\Phi) + B(\Phi) + C(\Phi,\phi,\kappa) - 2(\mu\kappa) D(\Phi,\phi,\kappa) + (\mu^2\kappa) E(\Phi,\phi,\kappa)\}/2\gamma^2). \tag{19}$$

In (19), $C(\Phi,\phi,\kappa) = z^T(G^{-1} - F^{-1})z$, $D(\Phi,\phi,\kappa) = (1 - \phi) \nu^T F^{-1} z$ and $E(\Phi,\phi,\kappa) = k - 2(k-1)\phi + (k-2)\phi^2 + \kappa(1 - \phi^2 \nu^T F^{-1}\nu$, where the $1 \times k$ vector $\nu^T = (1, 1-\phi, 1-\phi, \ldots, 1-\phi, 1)$, where the matrix G is given by (14) and (15), and where z is given by (16). The matrix F in (19) is a symmetric $k \times k$ band matrix of bandwidth 3, whose diagonal entries are $g_{tt} + \kappa(1 + \phi^2)$ for $t = 2, \ldots, k-1$ and $g_{tt} + \kappa$ for $t = 1$ or k, and whose nonzero off-diagonal entries are $g_{t,t+1} + \phi\kappa$ for $t = 1, \ldots, k-1$. On setting the derivatives of (19) with respect to ξ and γ equal to 0, the maximum likelihood estimates of these parameters, given Φ, ϕ and κ, may be easily found. Upon derivation of the variance of $\hat{\xi}$ under the composite model (10) and (18), and on using (6), $L_M(\Phi,\phi,\kappa)$, the marginal and approximate conditional likelihood, although a complicated function of Φ, ϕ and κ, may be easily expressed. Since the total number of parameters is small the maximum likelihood estimate and the maximum marginal likelihood estimate are both consistent and asymptotically unbiased, and will likely be close in value. Exact likelihoods, though they may be complicated expressions, may be obtained when (10) and (18) are replaced by general stationary autoregressive-moving average models. Likewise, the associated marginal and approximate conditional likelihoods may be derived.

The second approach is a two-step procedure. Under the fixed regression parameter model, the marginal or approximate conditional likelihood has a very simple form, given by (4) or (5). In the context of sampling on successive occasions with a first-order autoregressive model and simple random sampling, the marginal and approximate conditional likelihood is given by (17). Moreover, the value of the marginal likelihood for any given value of the parameters of Ω may be easily obtained on a direct application of the state-space model approach given by Harvey and Phillips (1979). Once the random coefficient model is used, for example the model of Blight and Scott (1973) in sampling on successive occasions, both the likelihood (full, marginal or approximate conditional) and the state-space models to apply the approach in this context, become much more complicated. In addition, model identification, for example (18) or a higher order process, is not straightforward. In view of the desire for simplicity, with perhaps only a small loss in efficiency, the following scheme may be suggested for the estimation of the parameters in Ω by using the marginal or conditional likelihood approach conditional on the occasion means $\mu_1, \ldots, \mu_k$(or conditional on β in the regression context). As the number of occasions k increases, so will the number of model parameters increase. In situations in which there are relatively short time series on individual units as in the case of sampling on successive occasions, the maximum likelihood estimates of the parameters in Ω may be biased and inconsistent. However, as Cruddas et al. (1989) have shown empirically for an autoregressive process of order one, the use of the marginal or approximate conditional likelihood to estimate the correlation parameters corrects this problem. Once estimates of the parameters of Ω have been obtained then estimates $\hat{\mu}_1, \ldots, \hat{\mu}_k$ of $\mu_1, \ldots, \mu_k$ may be obtained by the methods outlined by Harvey and Phillips (1979). Now, for example, under model (10) and (18) with the process in (18) replaced by a general ARMA process, the variance-covariance matrix of $\hat{\mu}_1, \ldots, \hat{\mu}_k$ is given by $\sigma^2 G^{-1} + \gamma^2\Gamma$. If $\sigma^2 G^{-1}$ is small compared to $\gamma^2\Gamma$, which may be the case when the sample sizes for the elementary surveys estimates are large, then $\hat{\mu}_1, \ldots, \hat{\mu}_k$ may be used with little loss in efficiency as the data to identify the process and estimate the parameters in Γ. Revised estimates of μ_t may be obtained using the estimated process.

4. COMPLEX SURVEYS

There are several ways in which one may proceed to analyze time series data from complex surveys. Each method that can be put forward will depend upon the sample information that is available.

If data are available at the micro level, then variance-covariance matrices based on the complex design can be computed for the elementary estimates for each rotation group. For the situation in which $\mu_1, \ldots, \mu_k$ are treated as fixed, a pseudo marginal likelihood is given by (4) and (9) with $\hat{x}_r$ and S_r replaced by their complex survey counterparts. A similar approach is taken, for example, by Roberts, Rao and Kumar (1987) in logistic regression analysis for complex surveys: obtain a likelihood or a set of likelihood equations and replace the usual statistics by their complex survey counterparts. For random model means, one option is to proceed with the fixed means analysis as the first step in the two-step estimation procedure described in section 3.3. Another option is to obtain the marginal likelihood under the random means model, for example the likelihood in

(19) and the marginal likelihood that may be derived from it. Then the statistics in this marginal likelihood are replaced by their complex surveys counterparts to obtain a pseudo marginal likelihood.

In many cases the data at the micro level will not be available. The estimation procedure then depends upon the data that are available. Two scenarios are considered here; many more could be formulated. In the first scenario, the sample covariances or correlations are not available, while in the second, they are.

Suppose that only the elementary estimates and their design effects are available. Let $\bar{y}_{t,r}$ be the estimate from rotation group G_r on occasion t based on a sample of size m_r. Let $\text{deff}_{t,r}$ be the design effect associated with $\bar{y}_{t,r}$. If σ^2/m_r is the variance of $\bar{y}_{t,r}$ under simple random sampling, then on appealing to the Central Limit Theorem,

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} \sim N(0,\sigma^2/m_r) \tag{20}$$

approximately. The modelling may proceed by assuming, within G_r, an ARMA-type process such as

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} = \phi(\bar{y}_{t-1,r} - \mu_{t-1})/(\text{deff}_{t-1,r})^{1/2} + \varepsilon_t, \tag{21}$$

where ε_t has constant variance. This may be easily cast into the framework of model (1), where the data vector y contains data of the form $\bar{y}_{t,r}/(\text{deff}_{t,r})^{1/2}$, where β is $(\mu_1, \mu_2, \ldots, \mu_k)^T$, and where X contains entries of the form $1/(\text{deff}_{t-1,r})^{1/2}$. The marginal likelihood, obtained as a special case of (5) or (6), may be evaluated using the state space models of Harvey and Phillips (1979) as noted in section 2. Marginal and approximate conditional likelihood estimation is desirable under the model given by (20) and (21). The estimate of ϕ in this case is based on the variation between elementary estimates within each rotation group; the variation within elementary estimates is not available. The length of time a rotation group remains in the sample is short so that the problems of bias and inconsistency in the maximum likelihood estimates will be applicable here.

If model (21) is combined with, for example, model (10), then the two-step procedure, as outlined in section 3.3 may be used to estimate the autoregressive parameter in (10).

For the second scenario, suppose that the survey estimates of the mean, say $\bar{y}_t$, are available for each occasion $t = 1, \ldots, k$. Also, the matrix, say S, of variances and covariances of the surveys estimates is available. In this situation a pseudo marginal likelihood can be obtained from (6). As in Binder and Dick (1989), among several others, the $\bar{y}_t$'s may be modelled by

$$\bar{y}_t = \mu_t + e_t, \tag{22}$$

where e_t is the survey error at time t with variance-covariance matrix estimated by S. The means on each occasion, μ_t for occasion t, follow an ARMA process. Since this is a special case of the random coefficients regression model, the appropriate marginal likelihood may be obtained through (6). Since S is available, an estimate of Ω, the correlation matrix of the survey error, may be easily obtained. An estimate of $\kappa = \sigma^2/\gamma^2$, may also be obtained. From assumptions which lead to the marginal likelihood in (6), it is necessary to assume that e_t in (22) is a stationary random variable. Then an estimate of σ^2 is the average of the diagonal elements in S. If γ^2 is the variance of the μ's then the variation between $\bar{y}_t$, $t = 1, \ldots, k$ provides an estimate of $\sigma^2 + \gamma^2$. From these two estimates, an estimate of κ may be obtained. Under model (22), X in (6) is the $k \times k$ identity matrix, while W is a $k \times 1$ column vector of 1's. Then the pseudo marginal likelihood for Γ (pseudo since κ and Ω have been replaced by their estimates) is given by (6) with the appropriate substitutions. If k, the number of occasions, is relatively large in comparison to the number of parameters in Γ, then the marginal and approximate conditional likelihood estimates should be similar to the maximum likelihood estimator. For ease of computation, it seems that the full likelihood approach using the state space models as outlined by Binder and Dick (1989, section 3) appears to be the simplest approach to use in this situation.

5. DISCUSSION

Marginal and approximate conditional likelihood techniques can be applied in a variety of situations for sampling on successive occasions. Since marginal likelihood methods show substantial improvements over maximum likelihood estimation when the number of nuisance parameters is large, use of these likelihood techniques may be recommended for use in the fixed means model such as (7) or in the random means model using a two-step estimation procedure as outlined in section 3.3. State space models may be easily applied in these situations to evaluate the marginal likelihood. In other situations where the number of nuisance parameters is small, such as the random means model outlined in (22), the use of the full likelihood is preferred.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Bellhouse, D.R. (1978). Marginal Likelihoods for distributed lag models. *Statist. Hefte* **19:** 2-14.

Bellhouse, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *J. Statist. Plan. Inf.* **21:** 69-74.

Bellhouse, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, to appear.

Binder, D.A. and Hidiroglou, M.A. (1988). Sampling in time. In: *Handbook of Statistics, Volume 6 (Sampling)*, P.R. Krishnaiah and C.R. Rao (eds.). Amsterdam: North-Holland, pp. 187-211.

Binder, D.A. and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology* **15:** 29-45.

Blight, B.J.N. and Scott, A.J. (1973). A stochastic model for repeated surveys. *J, Roy. Statist. Soc. (B)* **35:** 61-68.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. (B)* **49:** 1-39.

Cruddas, A.M., Reid, N., and Cox, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76:** 231-237.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Ann. Math. Statist.* **38:** 1456-1465.

Harvey, A.C. and Phillips, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika* **66:** 49-58.

Kalbflelsch, J.D. and Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. (B)* **32:** 175-208.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *J. Roy. Statist. Soc. (B)* **12:** 241-255.

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.* **20:** 335-375.

Roberts, G., Rao, J.N.K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74:** 1-12.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *J. Roy. Statist. Soc. (B)* **51:** 15-27.

PART 2

TIME SERIES ANALYSIS IN THE PRESENCE OF SURVEY ERROR

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

A TIME SERIES MODEL FOR ESTIMATING HOUSING PRICE INDEXES ADJUSTED FOR CHANGES IN QUALITY

D. Pfeffermann, L. Burck, and S. Ben-Tuvia[1]

ABSTRACT

The estimation of housing price indexes is based on the sale prices of homes sold in successive time periods. As such, the transactions recorded are not under control and they usually include homes of different quality in different time periods. The common approach to adjust for quality changes (e.g. in the computation of automobile price indexes) is to regress the sale prices against variables measuring quality. However, the housing price indexes are required separately for numerous cells with only few or even no transactions being recorded in many of these cells at the time that the indexes are calculated. In order to deal with this problem we propose the use of a dynamic linear model which accounts for the time series relationships between the cell regression coefficients and allows for contemporary correlations between coefficients operating in neighbouring cells. Modifications to ensure the robustness of the model and control its performance in periods of accelerated inflation are proposed. Empirical results illustrating the performance of the model in comparison to models which postulate fixed regression coefficients are presented using data on home prices in the city of Jerusalem for the years 1982-1989.

KEY WORDS: Hedonic Regression, Laspeyres Index, Robust Prediction, State Space Model.

1. INTRODUCTION

The consumer price index (CPI) is one of the most important and widely used economic series. It constitutes a major indicator of economic developments and often serves as a basis for salary and wage contracts as well as contracts in capital markets. Another important use of the CPI is to serve as a deflator for converting statistical series expressed in current prices to the same series expressed in constant prices of a given period.

Ideally, the CPI is intended to measure the effect of price changes on the budget required by consumers to maintain a given level of consumption. In practice, the index measures the percentage change over time in the expenditure required to consume a fixed "basket" of commodities and services. The items included in the basket and their relative weights are determined periodically on the basis of a family expenditure survey so that the basket represents the average consumption of the population to which the index refers.

In the present study we confine the discussion to the Laspeyres price index which is the index in common use. Denoting by P_{ko} and Q_{ko} the price and quantity of an item k in a base period and by P_{kt} the corresponding price of the same item in time period t, the Laspeyres index for time t is defined as

$$L_t = \sum_k P_{kt} Q_{ko} / \sum_k P_{ko} Q_{ko} = \sum_k \frac{P_{kt}}{P_{ko}} W_k \tag{1.1}$$

where the summation is over all the items ineluded in the basket and $W_k = P_{ko} Q_{ko} / \sum_k P_{ko} Q_{ko}$. Written in this manner, the index can be viewed as a weighted average of the price indexes $R_{kt} = (P_{kt}/P_{ko})$ of the goods and services included in the basket with the weights representing the relative expenditures of the corresponding items in the base period. Item k may itself be an aggregate of a number of sub-items in which case the index R_{kt} is again computed as a Laspeyres index of the sub-items composing the item k. This method is usually applied over several levels of aggregation, depending on the good or service under consideration.

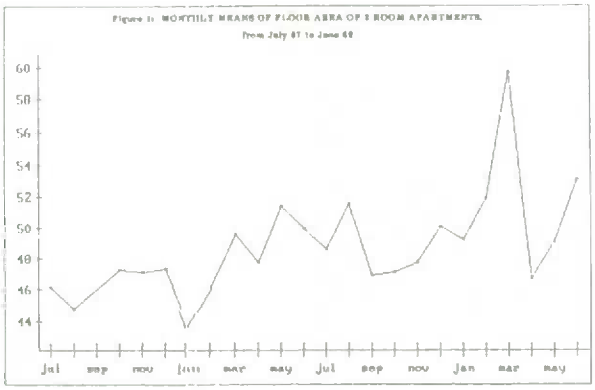
In order to assure that the index reflects only changes in prices of the goods and services and not other changes, it is imperative that the prices recorded in successive time periods will refer to the same or equivalent items. However, this requirement is frequently problematic. Some food and wear items are seasonal and are not available in every time period. Among durable goods new models come out which have different qualities from the models introduced in previous periods. This last is a perennial problem in calculating annual price indexes for road vehicles.

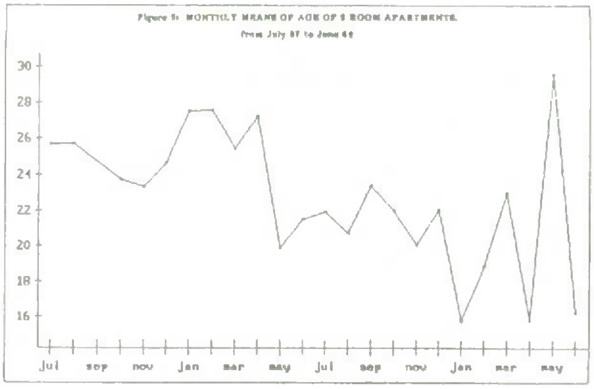
When computing housing price indexes (HPI), the changes in quality between adjacent time periods arise from the fact that the transactions performed in any two periods are not under control and they usually involve different types of housing. In Israel, where our empirical data come from, this problem is of particular concern because the aggregate HPI is a weighted average of HPI's computed in small cells classified by geographic units

[1] D. Pfeffermann, Hebrew University, Jerusalem, Israel 91905, L. Burck, Central Bureau of Statistics, Jerusalem, Israel 91130, S. Ben-Tuvia, Central Bureau of Statistics, Jerusalem, Israel 91130

(towns) and the home size (number of rooms). With time intervals of only one month, the actual number of transactions carried out and processed in time can be very low in many of these cells giving rise to large differences in quality.

As an illustration, we show below the monthly means of age and floor area of 2 room apartments in the city of Jerusalem for the months of July 1987 - June 1989. The number of transactions which these means are based on ranges from 5 to 69. (The numbers are particularly small in the last 3 months because most of the transactions are usually recorded only within 3 months after the HPI is first calculated and published.)





The problem of quality changes in the computation of price indexes has many facets and it had been widely discussed in the literature. See for example the books by Hofsten (1952) and Griliches (1971). (Both books share the same title - "Price Indexes and Quality Change"). Most of the studies in this area focus however on the computation of aggregate price indexes for durable goods so that the emphasis is on the ways by which to account for technical improvements and the addition of new features and not on changes in quality caused by small sample sizes. As Griliches notes, "most of the workers in this area, including myself, tried to get as large a cross section in any year as possible, not worrying too much about the overall comparability of any two cross sections" (Griliches, 1971, p. 7).

In the present article we consider this different aspect of the quality change problem namely, the change implied by the use of small samples which are not under control. We focus our attention to the computation of housing price indexes based on actual sale prices which prompted this study. In Israel (as in many other countries), the HPI is a major component of the CPI with a weight of about 15 percent. In addition, the HPI serves as an important economic indicator and is used for the linkage of contracts in construction and house rentals.

It should be emphasized that the use of actual sale prices (often referred to in the literature as the "Home Purchase Approach") is only one alternative for the computation of the HPI. In fact, there are at least four different such approaches with different countries adopting different methods at different times. Thus, while the Bureau of Labour Statistics in the U.S. used the home purchase approach until 1983 when it adopted a "Rental Equivalence Approach", the practice in New Zealand was to change from the rental equivalence approach to the home purchase approach. Castles (1987) provides an excellent review of the alternative approaches and summarizes the practices in over 130 countries.

Although we study the problem in the context of housing price indexes, the approach outlined in the present article can be applied after certain modifications to other price indexes of similar nature, e.g. the computation of price indexes of used cars. Furthermore, the model we use is a regression model with stochastic coefficients that can vary cross-sectionally and over time. Such a model has a large variety of applications in statistical and econometric studies.

The content of the article is as follows: in the next section we review the Hedonic regression approach for the adjustment of quality changes and describe its application in Israel pointing out the problems underlying its use. In section 3 we define the proposed model and discuss its properties. Estimation of the model parameters is considered in section 4. A modification to ensure the robustness of the model and control its performance in periods of accelerated inflation is proposed in section 5. Section 6 contains empirical results illustrating the important features of the model. We conclude the article with an outline for further analysis in section 7.

This article is of an expository nature and as such, the technical details have been reduced to the minimum necessary. The missing mathematical derivations can be obtained from the authors.

2. QUALITY ADJUSTMENT USING HEDONIC REGRESSIONS

The common procedure to adjust for changes in quality is by use of "hedonic" regression as originated from the works of Court (1939), Stone (1956), and Adelman and Griliches (1961). (The first and the third studies deal with the computation of automobile price indexes. The second study considers price indexes in the national accounts.)

The hedonic regression approach has two variants. In the first variant, the transaction costs corresponding to a given time period are regressed against quality measure variables (QMV). Using the estimated coefficients, a 'mean transaction cost' is estimated for each of the time periods by computing the fitted regression values at fixed 'average' values of the QMV. Calculating ratios of these means yields the desired indexes. In the second variant, the transaction costs of several time periods are regressed against the QMV and time dummy variables with the coefficients of the latter being interpreted as estimates of the pure price change. (The regression coefficients of the other variables are assumed fixed over the time periods considered.)

The rationale underlying the two approaches is that 'most' of the transaction cost variation can be explained by a relatively small number of QMV (referred to as "characteristics" in the hedonic context) with the other, omitted aspects of quality assumed to be uncorrelated with the included ones. The first variant allows the regression coefficients to vary over time whereas under the second variant the weights of the QMV are held fixed, postulating therefore that any change in the average prices between successive time periods is encompassed in the coefficients of the time dummy variables. Assuming that the separate regression equations used for the first variant include intercept terms, it can easily be seen that the combined model holding over the various time periods includes the model used under the second variant as a special case. The theoretical aspects of the use of these two approaches are discussed in Griliches (1971). (See also the discussion at the end of this section).

In Israel, the Central Bureau of Statistics (CBS) adopted a modified version of the second variant for the computation of the HPI's. Three QMV are used in the regression: Floor area (in square meters), Age (in years) and District (defined by one or two dummy variables depending on the size of the city).

The computations consist of three stages:

Stage 1: For each cell defined by city and number of rooms, with sufficient data, a multiplicative regression model is estimated every three months using the data available for the most recent six month period. The regression equation has the form

$$\mathrm{Log}\ Y_{tkj} = \alpha_0 + \alpha_{k1} \log F_{tkj} + \alpha_{k2} \log A_{tkj} + \alpha_{k3} D^{(1)}_{tkj} + \alpha_{k4} D^{(2)}_{tkj} + g_k(t) + \epsilon_{tkj} \qquad (2.1)$$

where Y_{tkj} is the cost of the j-th transaction in cell k during month t, F_{tkj}, A_{tkj}, $D^{(1)}_{tkj}$ and $D^{(2)}_{tkj}$ are the corresponding floor area, age and the two district indicator variables (only one indicator variable is considered in the small cities) and ϵ_{tkj} is a random disturbance assumed to have constant variance $\sigma^2_k = E(\epsilon^2_{tkj})$.

The time function $g^{(t)}_k$ is piecewise linear and it is defined for t=1 ... 6 as follows (t=6 represents the most recent month with data).

$$g_k(t)=\lambda_{k1}t_1 + \lambda_{k2}t_2 \quad ; \quad t_1 = \begin{cases} t & \text{if } t < 4 \\ 3.5 & \text{otherwise} \end{cases} , \quad t_2 = \begin{cases} 0 & \text{if } t < 4 \\ t-3.5 & \text{otherwise} \end{cases} \qquad (2.2)$$

The model defined by (2.1) and (2.2) is estimated using ordinary least squares (OLS) yielding preliminary estimates $(\hat{\lambda}_{k1}, \hat{\lambda}_{K2})$ with estimated variances $\{\hat{V}(\hat{\lambda}_{k1}), \hat{V}(\hat{\lambda}_{k2})\}$.

Stage 2: In stage 2 the estimates $(\hat{\lambda}_{k1}, \hat{\lambda}_{K2})$ are "shrinked" towards a common mean calculated from estimators obtained for neighbouring cells. The neighbouring cells used for the shrinkage process are all the cells pertaining to the same city if sufficient data is available or the cells pertaining to a group of cities otherwise. The shrinkage is carried out by considering the λ-coefficients operating in a given group of cells as exchangeable independent random variables such that

$$E(\lambda_{ka}) = \lambda_a; \quad E(\lambda_{ka} - \lambda_a)(\lambda_{\ell b} - \lambda_b) = \begin{cases} \delta^2_a & a = b,\ k = \ell \\ 0 & \text{otherwise} \end{cases} \quad a,b,=1,2 \qquad (2.3)$$

The modified, shrinked estimates are the empirical extended least square estimates (Pfeffermann and Nathan, 1981) defined as

$$\hat{\lambda}_{ka}(e) = G_K \hat{\lambda}_{Ka} + (1 - G_K)\hat{\lambda}_a(e) \quad ; \quad \hat{\lambda}_a(e) = \sum_K G_K \hat{\lambda}_{Ka} / \sum_K G_K \qquad (2.4)$$

where $G_K = \hat{\delta}^2_a / \{\hat{\delta}^2_a + \hat{V}(\hat{\lambda}_{Ka})\}$. The variances δ^2_a are estimated using the iterative procedure proposed by Pfeffermann and Nathan (1981) which is applied to all the cell estimates of all the groups, so that only one variance estimate is used in every quarter for each of the two λ coefficients. The λ-coefficients of cells with insufficient data to allow the computation of the OLS estimators are estimated by the corresponding means $\hat{\lambda}_a(e)$, $a=1,2$. For notational convenience we use below the symbols $\lambda_{Ka}(e)$ for all the cells regardless of data availability.

Stage 3: Using the model defined by (2.1) and (2.2), an HPI is estimated for each of the cells for a time span of 3 months. The index represents the average price increase between month 2 (the mid-point of the first quarter) and month 5 (the mid-point of the second quarter) and it is calculated as

$$\hat{R}_{K,5/2} = \hat{Y}_{5K\cdot}/\hat{Y}_{2K\cdot} = \exp\{1.5\ \hat{\lambda}_{K1}(e) + 1.5\ \hat{\lambda}_{K2}(e)\}$$

where $\hat{Y}_{tK\cdot}$ is the predicted (fitted) price at time t for given average values of the QMV. Notice that as a result of the use of a multiplicative relationship and the assumption of fixed coefficients during the six month period, the ratio $\hat{R}_{K,5/2}$ is independent of the choice of the average values of the QMV. Another noteworthy point is that under the assumption of normality for the error terms, the ratio $\hat{R}_{k,5/2}$ is biased as an estimator of $R_{k,5/2} = \{E(Y_{5K\cdot}) / E(Y_{2k\cdot})\}$ but the bias was found to have a negligible effect on the estimated MSE of the estimators and hence is ignored when constructing the index.

Having calculated the cell indexes, they are aggregated to form higher level indexes using appropriate cost weights obtained from the most recent family expenditure survey. Monthly indexes are calculated by interpolation using the corresponding changes in the index of "Inputs in Residential Building" as benchmarks. The monthly indexes are then incorporated in the CPI.

Due to late registry of some of the transactions and administrative delays in processing, data pertaining to a given month may become available up to three months later. Using the delayed data, the HPI is revised after 3 months, concurrent with the computation of the new HPI. However, the revised HPI, although more stable, is only of limited use.

DISCUSSION: The method described above has some clear shortcomings. The assumption that the marginal effects of the QMV remain fixed throughout a six month period and that the price changes are reflected solely in the time function goes against much of the index number literature and is at best a crude approximation. It implies under the multiplicative relationship (2.1) that the ratio between the expected prices of homes of different fixed qualities remains constant throughout the time period of six months. The housing market is an unstable market determined by negotiations between sellers and buyers which are affected by the concurrent state of the economy and as such, it seems much more appropriate to let the coefficients of the QMV to change over time. (The instability of econometric relationships is often argued in the literature, see for example the discussion in Cooley and Prescott, 1976). The particular choice of the time function although based on some empirical evidence in a particular year is clearly not flexible enough to account for the month to month changes in the prices of homes and not general enough to hold simultaneously in all the time periods and for all the different types of housing. Another limitation of the current procedure is the interpolation of the monthly indexes which is done in a rather ad-hoc manner.

The obvious reason for the use of this particular method by the CBS is the lack of sufficient data, even for the larger cells at the time that the HPI is calculated. While an attempt is made to borrow information from neighbouring cells, this does not solve the other problems listed above. It seems that a major source of information not exploited in the current procedure is the time series properties of the data. As it stands, data prior to the six months period under consideration are ignored when computing the current indexes despite the fact that these data pertain to the same cells and measure the same phenomenon. The model presented in the next section accounts for both the time serie and the cross-sectional relationships between the regression coefficients. By borrowing information from the past, the estimation of the indexes can be carried out on a monthly basis without the need to impose constant coefficients for the QMV or postulate a deterministic time function to represent the price changes which are the major limitations of the current procedure.

3. REGRESSION WITH COEFFICIENTS THAT VARY CROSS-SECTIONALLY AND OVER TIME

In what follows we denote by $\underset{\sim}{Y}_{tk}$ the $(n_{tk} \times 1)$ vector of observations on the dependent variable (logs of transaction costs in our case) pertaining to cell (domain) k at time t, $k=1\ldots K$, $t=1, 2, \ldots.$ We assume that $\underset{\sim}{Y}_{tk}$ is nonempty although as will become evident in section 4, having no observations in some of the cells at certain times causes no methodological difficulties. Let X_{tk} represent the corresponding model (design) matrix of explanatory variables (QMV in our case). The regression model in cell k is defined as

$$\underset{\sim}{Y}_{tk} = \underset{\sim}{1}_{ntk}\, \gamma_{tk} + X_{tk}\, \underset{\sim}{\beta}_{tk} + \underset{\sim}{\varepsilon}_{tk}\ ; \qquad E(\underset{\sim}{\varepsilon}_{tk}) = \underset{\sim}{0},\ E(\underset{\sim}{\varepsilon}_{tk}\, \underset{\sim}{\varepsilon}'_{tk}) = \sigma_k^2\, I_{ntk} \tag{3.1}$$

where $\underset{\sim}{1}_{ntk}$ and I_{ntk} are correspondingly the unit vector and identity matrix of order n_{tk}. The notable feature of equation (3.1) is that the coefficients γ_{tk} and $\underset{\sim}{\beta}_{tk}$ are allowed to vary cross-sectionally and over time. The following equations specify the variation of the coefficients over time,

$$\gamma_{tk} = \gamma_{t-1,k} + s_{t-1,k} + n_{\gamma tk} \quad ; \qquad E(n_{\gamma tk}) = 0,\ E(n_{\gamma tk})^2 = \delta_\gamma^2$$

$$s_{tk} = s_{t-1,k} + n_{stk} \quad ; \qquad E(n_{stk}) = 0,\ E(n_{stk}^2) = \delta_s^2 \tag{3.2}$$

$$\underset{\sim}{\beta}_{tk} = \underset{\sim}{\beta}_{t-1,k} + \underset{\sim}{n}_{\beta tk};\ E(\underset{\sim}{n}_{\beta tk}) = \underset{\sim}{0},\ E(\underset{\sim}{n}_{\beta tk}\ \underset{\sim}{n}'_{\beta tk}) = \Delta_\beta,\ E(\underset{\sim}{n}_{\beta tk}\ n_{\gamma tk}) = {}_{\beta}\underset{\sim}{\delta}_{\gamma}$$

It is assumed also that n_{stk} is uncorrelated with $(n_{\gamma tk}, \underset{\sim}{n}_{\beta tk})$ and that all the serial correlations are equal to zero.

The implication of equations (3.2) is that they define a local approximation to a linear trend for the intercept term and a random walk model for the other coefficients. Since the explanatory variables are usually correlated, the changes in the values of the various coefficients may likewise be correlated which is accomodated by allowing for a general V-C matrix Δ_β (allowing in particular different residual variances for difference coefficients) and a general covariance vector ${}_{\beta}\underset{\sim}{\delta}_{\gamma}$.

A simple way to account for the cross-sectional relationships between the regression coefficients is by allowing for non-zero correlations between the corresponding residual terms of the equations (3.2). However, even with a small number of cells, one has to impose a certain structure on these correlations if the number of unknown model parameters is to be kept at a manageable level. One possibility which seems particularly useful in the case of a small number of cells is to assume constant correlations between the residual terms operating in different cells. Denoting $\underset{\sim}{n}'_{tk} = (n_{\gamma tk}, n_{stk}, \underset{\sim}{n}_{\beta tk})$, this assumption can be formulated as

$$E(\underset{\sim}{n}_{tk}\ \underset{\sim}{n}'_{t\ell}) = \Delta\emptyset\ ,\ k \neq \ell \tag{3.3}$$

where Δ is diagonal with δ_γ^2, δ_s^2 and the diagonal elements of Δ_β on the main diagonal and $\emptyset$ is another diagonal matrix with all its elements being inside the interval (-1,1). The diagonal elements of $\emptyset$ define the correlations between residual terms pertaining to different cells.

Another possibility applicable in the case where a "distance" can be measured between the various cells (like in the present study where the cells are defined by the number of rooms) is to postulate that the correlations between the residual terms decay as the distance between the cells increases. This assumption can be formulated as

$$E(\underset{\sim}{n}_{tk}\ \underset{\sim}{n}'_{t\ell}) = \Delta\emptyset f(k,\ell);\ k \neq \ell \tag{3.4}$$

where $f(k,\ell)$ is a monotonic decreasing function of the distances $D(k,\ell)$. Equation (3.3) is an obvious special case of (3.4).

DISCUSSION: The use of stochastic regression coefficients to account for time and/or cross-sectional variation of the regression coefficients is common in the statistical and econometric literature. Johnson (1977, 1980) provides an annotated bibliography of over 150 articles which consider models of this kind. Our model extends on previous models by postulating local linear trends for the intercept coefficients and by imposing a structure on the cross-sectional correlations. Cooley and Prescott (1976) and LaMotte and McWhorter (1977) assume that all the regression coefficients in their model follow a random walk, Rosenberg (1973a) assumes autoregressive relationships where as Hsiao (1974) and Swamy and Mehta (1977) assume that the coefficient realizations can be factorized into a common mean and independent error components which account for the time and the cross-sectional variation. For a review of these and the many other studies on regression with stochastic coefficients see the discussions in Rosenberg (1973b), Maddala (1977, Chapter 7), Dielman (1983) and Pfeffermann and Smith (1985).

The reasons for permitting the regression coefficients to vary over time have already been discussed at the end of section 2. The random walk model implies that the coefficients drift gradually away from their initial value with no inherent tendency to return to a mean value. This kind of model appeals to us as being appropriate for fitting the home purchase costs. It has the further advantage of being parsimonious in terms of the number of unknown parameters which is very important in view of the already large number of parameters included in the equations (3.1) - (3.3).

The particular choice of the model for the intercept term was dictated by the relatively high monthly inflation rates in Israel, fluctuating around 1.5 percent in the last two years. This means that we would expect the log of the prices of given homes (the dependent variable in our model) to grow approximately linearly over time which, for fixed values of the other regression coefficients would imply an approximately linear trend for the intercept term as defined by the first two equations of (3.2).

The model defined by (3.1) - (3.3) overcomes the limitations of the CBS procedure discussed at the end of section 2. The weights assigned to the various QMV are no longer fixed over time and the deterministic time function (2.2) is replaced by a more flexible and time adapting trend function. The estimators derived for any given cell are strengthened by borrowing information from both neighbouring cells and from past data. The amount of information borrowed is determined by the nearness of the vectors of coefficients (cross-sectionally and over time) as detected by the estimators of the model variances and covariances (see section 4 for details).

4. MODEL ESTIMATION

4.1 Model Representation in State Space Form

In what follows we use the following notation: we define $\underset{\sim}{Y}'_t = (\underset{\sim}{Y}'_{t1} \cdots \underset{\sim}{Y}'_{tK})$ to represent the vector of observations at time t of length $n_t = \sum_{k=1}^{K} n_{tk}$ and $\underset{\sim}{\varepsilon}'_t = (\underset{\sim}{\varepsilon}'_{t1} \cdots \underset{\sim}{\varepsilon}'_{tK})$ to represent the corresponding vector of residuals. Let $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, X_{tk}]$ where $\underset{\sim}{0}_{ntk}$ is the null vector of length n_{tk} and let Z_t be the block diagonal matrix with Z_{tk} comprising the k-th block. The matrix Z_t is of order $n_t \times (K \times m)$ where m denotes the number of columns in each of the matrices Z_{tk}. Define $\underset{\sim}{\alpha}'_{tK} = (\gamma_{tK}, s_{tK}, \underset{\sim}{\beta}'_{tk})$ to represent the regression coefficients corresponding to cell k and let $\underset{\sim}{\alpha}'_t = (\underset{\sim}{\alpha}'_{t1} \cdots \underset{\sim}{\alpha}'_{tK})$.

Using the above notation, the set by equations defined by (3.1) can be written compactly as

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\varepsilon}_t \; ; \qquad E(\underset{\sim}{\varepsilon}_t) = \underset{\sim}{0} \quad , \; E(\underset{\sim}{\varepsilon}_t \underset{\sim}{\varepsilon}_t') = \Sigma_t \tag{4.1}$$

where $\Sigma_t = \mathrm{Diag}\,[\sigma_1^2 1'_{nt1} \cdots \sigma_K^2 1'_{ntK}]$.

Let $T^{\star} = \begin{bmatrix} 1,1 & 0 \\ 0,1 & \\ 0 & I_{m-2} \end{bmatrix}$ be a block diagonal matrix of order mxm where I_{m-2} is the identity matrix of order (m-2) and define $T = I_K \otimes T^{\star}$ where $\otimes$ denotes the kronecker product.

The system of equations defined by (3.2) and (3.3) can be written compactly as

$$\underset{\sim}{\alpha}_t = T\underset{\sim}{\alpha}_{t-1} + \underset{\sim}{\eta}_t \quad ; \qquad E\,(\underset{\sim}{\eta}_t) = \underset{\sim}{0} \quad , \qquad E\,(\underset{\sim}{\eta}_t \underset{\sim}{\eta}'_t) = \Lambda \tag{4.2}$$

where $\underset{\sim}{\eta}'_t = (\underset{\sim}{\eta}'_{t1} \cdots \underset{\sim}{\eta}'_{tK})$ and $\Lambda = [\Lambda_{k\ell},\ k,\ell = 1 \ldots K]$ with

$$\Lambda_{kk} = E(\underset{\sim}{\eta}_{tk}\, \underset{\sim}{\eta}'_{tk}) = \begin{bmatrix} \delta^2_{\gamma} & 0 & {}_{\beta}\delta'_{\gamma} \\ 0 & \delta^2_{s} & \underset{\sim}{0}' \\ {}_{\beta}\delta_{\gamma} & \underset{\sim}{0} & \Delta_{\beta} \end{bmatrix} \quad \text{and} \quad \Lambda_{k\ell} = E(\underset{\sim}{\eta}_{tk}\, \underset{\sim}{\eta}'_{t\ell}) = \Delta\emptyset,\ k \neq \ell \; .$$

(The matrices Δ_{kk} and $\Delta_{k\ell}$ are of order mXm).

Equations (4.1) and (4.2) conform to the classical state-space model formulation — Harvey, 1984 with (4.1) representing the observations equation and (4.2) the system equation. The advantage of restructuring the model in a state space form is that the vectors $\underset{\sim}{\alpha}_t$ can be estimated then most conveniently by use of the Kalman filter. We describe the basic steps of the filter in the next section.

4.2 Model Estimation by Means of the Kalman Filter

In this section we assume that the V-C matrices Σ_t and Λ are known. Estimation of the unknown elements of these matrices is considered in Section 4.3. The Kalman filter consists of a set of recursive equations which define how to update current and past estimators of the system state vectors (the model regression coefficients $\underset{\sim}{\alpha}_t$ in our case) and how to predict future vectors every time that new data become available. In addition, the filter provides the V-C matrices of the various estimators and predictors. The theory of the Kalman filter is developed in numerous publications, (see e.g. Anderson and Moore, 1979 and Meinhold and Singpurwalla, 1983), and so we only present here the basic equations.

Let $\hat{\underset{\sim}{\alpha}}_{t-1}$ be the best linear unbiased predictor (blup) of $\underset{\sim}{\alpha}_{t-1}$ based on all the data observed up to time (t-1). Since $\hat{\underset{\sim}{\alpha}}_{t-1}$ is blup for $\underset{\sim}{\alpha}_{t-1}$, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}_{t-1}$ is the blup of $\underset{\sim}{\alpha}_t$ at time (t-1). Furthermore, if $P_{t-1} = E\,(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})'$ is the V-C matrix of the prediction errors at time (t-1), $P_{t|t-1} = TP_{t-1}T' + \Lambda$ is the V-C matrix of the prediction errors $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$. (Follows straightforwardly from 4.2).

When a new vector of observations becomes available, the predictor of $\underset{\sim}{\alpha}_t$ and the V-C matrix P_{t-1} are updated according to the formulae

$$\hat{\underset{\sim}{\alpha}}_t = \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1} (\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$$

$$P_t = (I - P_{t|t-1} Z_t' F_t^{-1} Z_t) P_{t|t-1} \qquad (4.3)$$

where $\hat{\underset{\sim}{Y}}_{t|t-1} = Z_t \hat{\underset{\sim}{\alpha}}_{t|t-1}$ is the blup of $\underset{\sim}{Y}_t$ at time (t-1) so that $\underset{\sim}{e}_t = (\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$ is the vector of innovations with V-C matrix $F_t = (Z_t P_{t|t-1} Z_t' + \Sigma_t)$.

The new data observed at time t can be used also for the updating (smoothing) of past estimators. Denoting by t* the most recent month with observations, the smoothing is carried out using the equations.

$$\hat{\underset{\sim}{\alpha}}_{t|t\star} = \hat{\underset{\sim}{\alpha}}_t + P_t T' P^{-1}_{t+1|t} (\hat{\underset{\sim}{\alpha}}_{t+1|t\star} - T \hat{\underset{\sim}{\alpha}}_t)$$

$$P_{t|t\star} = P_t + P_t T' P^{-1}_{t+1|t} (P_{t+1|t\star} - P_{t+1|t}) P^{-1}_{t+1|t} T P_t \; ; \; t=2,\ 3,\ \ldots\ t\star \qquad (4.4)$$

where $P_{t|t\star}$ is the V-C matrix of the prediction errors $(\hat{\underset{\sim}{\alpha}}_{t|t\star} - \underset{\sim}{\alpha}_t)$. Notice that $\hat{\underset{\sim}{\alpha}}_{t\star|t\star} = \hat{\underset{\sim}{\alpha}}_{t\star}$ and $P_{t\star|t\star} = P_{t\star}$ which defines the starting values for the smoothing equations.

When applying the model for the estimation of the HPI in a given month t, one needs to estimate the vectors $\underset{\sim}{\alpha}_t$ and $\underset{\sim}{\alpha}_{t-1}$. In order to estimate the variance of the estimated index it is necessary to estimate the covariance matrix between the estimators $\hat{\underset{\sim}{\alpha}}_t$ and $\hat{\underset{\sim}{\alpha}}_{t-1|t}$. The covariance matrix has the following form,

$$E(\hat{\underset{\sim}{\alpha}}_t - \underset{\sim}{\alpha}_t)(\hat{\underset{\sim}{\alpha}}_{t-1|t} - \underset{\sim}{\alpha}_{t-1})' = (I - P_{t|t-1} Z_t' F_t^{-1} Z_t) T P_{t-1} \qquad (4.5)$$

4.3 Estimation of the V-C matrices and Initialization of the Filter

The actual application of the Kalman Filter requires the estimation of the unknown elements of the matrices Σ_t and Λ and the initialization of the filter, that is, the estimation of the vector $\underset{\sim}{\alpha}_o$ and the corresponding V-C matrix P_o of the estimation errors. In this section we describe briefly the estimation methods used in the present study.

The unknown model parameters have been estimated using maximum likelihood theory. Assuming a normal distribution for the residual terms $\underset{\sim}{\varepsilon}_t$ and $\underset{\sim}{\eta}_t$ and a diffuse prior distribution for $\underset{\sim}{\alpha}_o$, the log likelihood function for the observations $\underset{\sim}{Y}_3 \ldots \underset{\sim}{Y}_t$ conditional on $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ can be formulated as

$$L(\underset{\sim}{\lambda}) = \text{constant} \;\; - \frac{1}{2} \sum_{t=3}^{T} (\log |F_t| + \underset{\sim}{e}_t' F_t^{-1} \underset{\sim}{e}_t) \qquad (4.6)$$

where $\underset{\sim}{\lambda}$ contains the unknown model variances and covariances written in a vector form. The expression (4.6) follows by using the prediction error decomposition, see Schweppe (1965) and Harvey (1981) for details. For given matrices Σ_t and Λ, the innovations $\underset{\sim}{e}_t$ and the V-C matrices F_t are obtained by application of the Kalman filter equations (4.3).

The computation of the likelihood function requires the initialization of the Kalman filter which was carried out using the approach proposed by Harvey and Phillips (1979). By this approach, the assumption of a diffuse prior for $\underset{\sim}{\alpha}_o$ is actualized by initializing the filter at time t=0 with $\underset{\sim}{\alpha}_o = \underset{\sim}{0}$ and $P_o = N \times I$ where N is a large finite number and I is the identity matrix of the appropriate order.

Maximization of the likelihood function (4.6) was implemented using the method of scoring with a variable step length. Let $\underset{\sim}{\lambda}_{(o)}$ define initial estimates of the unknown elements in $\underset{\sim}{\lambda}$. The method of scoring consists of solving iteratively the set of equations

$$\underset{\sim}{\lambda}_{(i)} = \underset{\sim}{\lambda}_{(i-1)} + r_i \{I[\underset{\sim}{\lambda}_{(i-1)}]\}^{-1} g[\underset{\sim}{\lambda}_{(i-1)}] \qquad (4.7)$$

where $\underset{\sim}{\lambda}_{(i-1)}$ is the estimator of $\underset{\sim}{\lambda}$ as obtained in the (i-1)th iteraction, $I[\underset{\sim}{\lambda}_{(i-1)}]$ is the information matrix evaluated at $\underset{\sim}{\lambda}_{i-1}$ and $g[\underset{\sim}{\lambda}_{(i-1)}]$ is the gradient of the log likelihood evaluated at $\underset{\sim}{\lambda}_{i-1}$. The coefficient r_i is the variable step length introduced to guarantee that $L[\underset{\sim}{\lambda}_{(i)}] \geq L[\underset{\sim}{\lambda}_{(i-1)}]$ in every iteration. The value of r_i was determined by a grid search procedure. The formulae for the k-th element of the gradiant vector and the kl-th element of the information matrix are given in Watson and Engle (1983).

Having estimated the model variances and covariances, they can be substituted for the true parameters in the Kalman filter equations (4.3) - (4.5) to yield the estimators of the regression coefficients and the V-C matrices. Notice that the estimated V-C matrices ignore the extra variability induced by the need to estimate the unknown elements contained in $\underset{\sim}{\lambda}$. Ansley and Kohn (1986) propose correction factors of order 1/t* to account for this extra variation in state space modelling.

A computer program which implements the methods described in this section for the estimation of the Kalman filter has been written using the procedure PROC-IML of the SAS system.

5. MODIFICATIONS TO PROTECT AGAINST MODEL BREAKDOWNS

5.1 Description of the problem and proposed modifications

The use of a model for calculating the HPI is inevitable in view of the quality change problem. It raises the question however of how to protect against possible model breakdowns. Testing the model every time that new data become available is not practical, requiring instead the development of a "built-in mechanism" which will secure the robustness of the indexes when the model fails to hold.

This problem is of particular concern in months where the prices show an unexpected jump. In Israel, for example, the currency is occasionallly devaluated in rates of up to 10 percent. While the devaluations are usually accompanied by strict price policies which attempt to freeze the old prices, these policies have little effect on home purchase prices which are determined by direct negotiations between buyers and sellers and hence are not under control. On the other hand, the model proposed in section (3) uses past relationships between prices and qualities to strenghten the estimation of current relationships and as such, it will adjust itself to such sudden changes only after a certain lag.

In order to deal with this problem we propose to modify the regression estimators derived in the various time periods so that they satisfy certain linear constraints obtained by equating aggregate means of the raw data with their expected values under the model. More precisely, we propose to augment the model equations (3.1) by linear constraints of the form

$$\sum_k W_{tk}^{(i)} (n_{tk}\gamma_{tk} + 1_{ntk}{}'X_{tk}\underset{-}{\beta}_{tk}) = \sum_k W_{tk}^{(i)} \sum_j Y_{tkj} \qquad \begin{matrix} i=1,2\cdots I(t) \\ t=1\cdots T \end{matrix} \qquad (5.1)$$

Where the coefficients $\{W_{tk}^{(i)}\}$ are fixed weights standardized to satisfy $\sum_k n_{tk} W_{tk}^{(i)} = 1$. It is important to emphasize that the constraints (5.1) do not represent external information about possible values of the regression coefficients. Rather, they serve as a control system to guarantee that the model estimators adjust themselves more rapidly to sudden changes in the behavior of the regression coefficients. As such, the variances of the modified regression estimators are slightly larger than the variances of the optimal estimators under the model. Obviously, when no such changes occur and the variances of the aggregate means are sufficiently small, one would expect the constraints to be satisfied approximately without imposing them explicitly. Ideally, one would like to incorporate several separate constraints in each time period but it is imperative that the variances of the corresponding aggregate means will be small enough to ensure that the modifications are indeed necessary and do not interfere with the random fluctuation of the raw data.

Examples of aggregate means which can be used in the case of the home purchase data include i) averaging separately over all the data included in cells with a large number of transactions, ii) averaging separately over combined cells of a given number of rooms, iii) averaging over cells with different number of rooms, e.g. over all the data pertaining to a given city. Notice that in view of the correlations between the regression coefficients operating in the various cells, a constraint applied to a sub-set of the cells will modify the regression estimates of all the cells. Battese, Harter and Fuller (1988) propose a similar kind of modification in the context of small area estimation.

5.2 Robust Estimation Using the Augmented Equations

In section 5.1 we proposed to amend the model equations (3.1) by imposing the set of constraints (5.1) thereby securing the robustness of the regression estimators against sudden drifts in the values of the coefficients. Computationally, this could be implemented most conveniently by augmenting the vectors $\underset{-}{Y}_t$ of equation (4.1) by the scalars $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ [$i=1,2\cdots I(t)$ indexes the number of constraints in time t], augmenting the matrices Z_t by the corresponding row vectors $(w_{t1}^{(i)} 1'_{nt1} Z_{t1} \cdots w_{tK}^{(i)} 1'_{ntK} Z_{tK})$ and setting the respective variances of the residual terms to zero. The augmented set of equations, together with (4.2), form a pseudo state-pace model which could be estimated using the Kalman filter equations (4.3). Notice that the pseudo V-C matrix $\Sigma_t^{(P)}$ of the augmented residual vector is no longer positive definite (the last $I(t)$ rows and columns of $\Sigma_t^{(P)}$ consist of zeroes) but this does not imply computational difficulties.

The drawback of applying the Kalman filter to the pseudo model is that the V-C matrices of the regression estimators fail to account for the actual variability of the aggregate means of the raw data. Although it was argued in section 5.1 that this variability could be ignored when the means are based on sufficiently large numbers of transactions, a better and more robust procedure would be to amend the formula for the updating of the V-C matrix P_t (equation 4.3) so that the variances and covariances of the random variables $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ will be taken into account. Let $\underset{\sim}{Y}_t^{(A)}$ and $Z_t^{(A)}$ represent the augmented Y vector and Z matrix at time t and denote by $\Sigma_t^{(A)}$ the actual V-C matrix of the residual terms $[\underset{\sim}{Y}^{(A)} - Z_t^{(A)} \underset{\sim}{\alpha}_t]$. The matrix $\Sigma_t^{(A)}$ is of order $[n_t + I(t)]$ with Σ_t in the first n_t rows and columns and the variances and covariances of the means $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ among themselves and with the vector $\underset{\sim}{Y}_t$ in the remaining rows and columns. Denoting by $\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}$ the robust predictor of $\underset{\sim}{\alpha}_{t-1}$ as obtained at time $(t-1)$ using the pseudo model and by $P_{t-1}^{(A)}$ the actual V-C matrix of the errors $(\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_{t-1})$, the modified state estimator at time t is obtained as

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + P_{t|t-1}^{(A)} Z_t^{(A)'} (F_t^{(P)})^{-1} [\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}] \tag{5.2}$$

where $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)} T' + \Lambda)$ and $F_t^{(P)} = [Z_t^{(A)} P_{t|t-1}^{(A)} Z_t^{(A)'} + \Sigma_t^{(P)}]$. (Compare with 4.3). It can be shown that the actual V-C matrix $P_t^{(A)}$ of the errors $(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t)$ satisfies the recursive equation

$$P_t^{(A)} = [I - K_t^{(P)} Z_t^{(A)}] \, P_{t|t-1}^{(A)} + K_t^{(P)} [\Sigma_t^{(A)} - \Sigma_t^{(P)}] K_t^{(P)'} \tag{5.3}$$

where $K_t^{(P)} = P_{t|t-1}^{(A)} Z_t^{(A)'} (F_t^{(P)})^{-1}$ is the pseudo Kalman gain. The first expression on the right hand side of (5.3) corresponds to the usual updating formula of the Kalman filter [compare with (4.3)]. The second expression is a correction factor which accounts for the actual variances and covariances of the means $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$, not taken into account in the first expression.

The amended Kalman filter defined by the equations (5.2) and (5.3) produces the robust predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ instead of the optimal model dependent predictors but uses the correct V-C matrices under the model. Thus, this filter can be used for the routine estimation of the vectors of coefficients and when the model holds it will give similar results to those obtained under the optimal filter. In periods where the model fails to hold, the updating formula (5.3) could be incorrect (depending on the particular model failures) but the predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ will nonetheless satisfy the linear constraints (5.1). The smoothing equations (4.4) and the V-C matrix in (4.5) can be modified to the case of using the robust predictors in a similar way.

6. EMPIRICAL RESULTS

In order to confirm and illustrate the appropriateness of the model to the home purchase prices in Israel, we fitted the model separately to the five cells in the city of Jerusalem using the data observed for the transactions performed during the period September 1982 - June 1988. The cells are defined by the number of rooms - ranging from 1 to 5. For time and other technical reasons we have not yet run the model incorporating simultaneously data from different cells so that the model uses only the time series relationships between the cell regression coefficients as defined by the equation (3.2). Since we only used data from one cell in each run we have also not incorporated the modifications discussed in section 5. A comprehensive computer program which fits the full model defined by the equations (3.1) - (3.3) using the estimation methods described in section 4 and incorporating the modifications of section 5 is now in a test process and will be available to interested readers upon request. (The raw data may likewise be provided).

The models fitted to the five cells were found to be generally consistent in terms of the significance of the model variance estimators. Thus, except for the case of 5 room apartments, the variance $\hat{\delta}_s^2$ of the slope coefficient was found to be insignificant, implying in turn a random walk model for the intercept coefficient since the initial slope coefficient was set to zero. For 5 room apartments, $\hat{\delta}_s^2 = 4 \times 10^{-4}$ which is very low although significant at the 0.5 percent level. Likewise, except for the case of 4 room apartments, the variances of the intercept and the other four QMV defined by the equation (2.1) were all found to be highly significant supporting our initial conjecture that the regression coefficients change stochastically over time. In the case of 4 room apartments floor, the variance of the intercept coefficient is again highly significant and the variance of the floor area slope coefficient is significant at the 10 percent level but the remaining variance estimators are not significant.

(In section 3, we suggested that in view of the relatively high and approximately constant monthly inflation rates in Israel, the intercept coefficients could be growing linearly. The discussion assumed however that the other coefficients are constant over time which is clearly not the case. It seems also that the HPI is much more variable compared to the overall consumer price index.)

In the remainder of this section we show several graphs illustrating the performance of the model in the case of 2 room apartments. The restriction to 2 room apartments is merely for space reasons and the results obtained for the other cells are generally very similar. We use the following definitions

Y^*_{tj} - the log of the sale price of apartment j in month t, $j=1 \ldots n_t$, $t=1,2 \ldots T$

$\underset{\sim}{x}_{tj}$ - the QMV corresponding to apartment j in month t. The QMV are the intercept and the four variables specified by equation (2.1) [excluding the time function $g_2(t)$]

$\hat{\underset{\sim}{\alpha}}{}^{OLS}_t$ - the OLS estimators of the QMV coefficients based on the transactions performed during month t.

$\hat{\underset{\sim}{\alpha}}{}^{F}_t$ - the filtered estimators of the QMV coefficients based on the transactions performed up to and including month t (equation 4.3)

$\hat{\underset{\sim}{\alpha}}{}^{S}_t$ - the smoothed estimators based on all the transactions performed in all the months (equation 4.4)

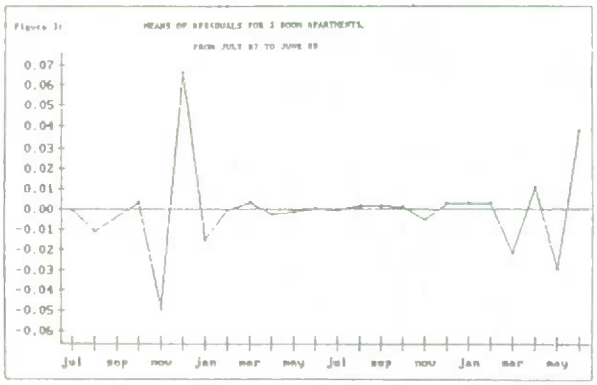
$\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}{}^{F}_{t-1}$ - the predicted values of the QMV coefficients one step ahead

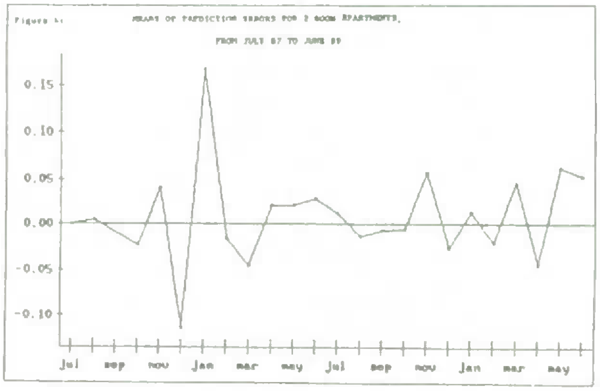
$e_{tj} = (Y^*_{tj} - \underset{\sim}{x}'_{tj}\, \hat{\underset{\sim}{\alpha}}{}^{F}_t)$ - the residual observed for transaction j in month t

$m_t = \sum_{j=1}^{n_t} e_{tj}/n_t$ and $mse_t = \sum_{j=1}^{n_t} e^2_{tj}/n_t$ - the monthly means and MSE's of the residuals

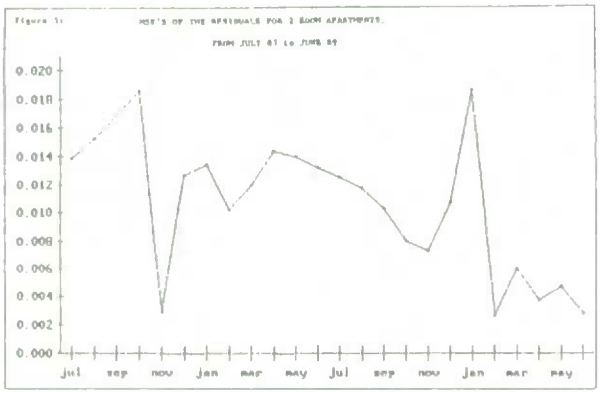
$e_{ptj} = (Y^*_{tj} - \underset{\sim}{x}'_{tj}\, \hat{\underset{\sim}{\alpha}}_{t|t-1})$ - the prediction error associated with transaction (tj)

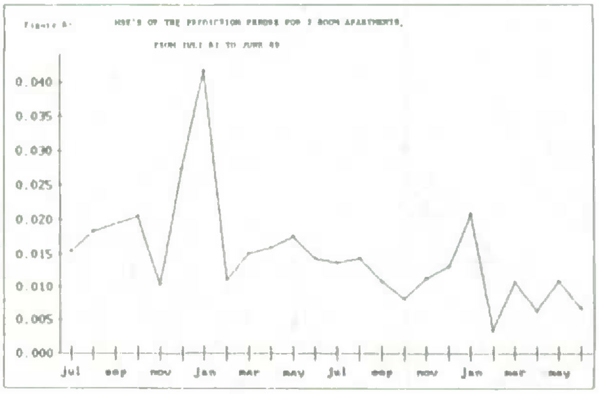
$m_{pt} = \sum_{j=1}^{n_t} e_{ptj}/n_t$ and $mse_{pt} = \sum_{j=1}^{n_t} e^2_{ptj}/n_t$ - the monthly means and MSE's of the prediction errors.



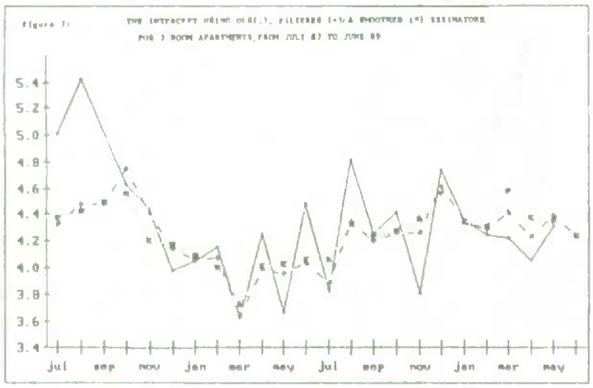


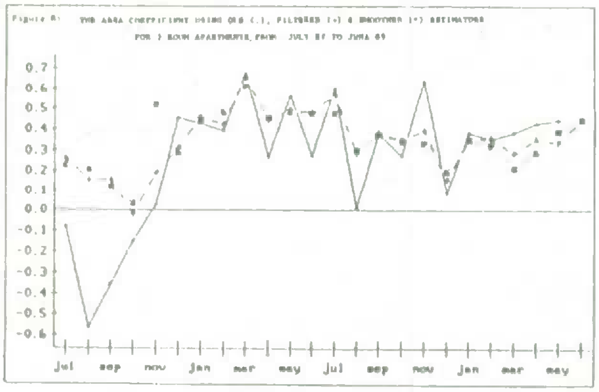
Figures 3 and 4 plot the monthly means of the residuals and the prediction errors for the months of July 87 - June 89. Figures 5 and 6 plot the corresponding MSE's. Notice that the last 12 months' data were not used for the estimation of the model variances. As could be expected, the prediction errors are more variable than the residuals but there is nothing in the four Figures to indicate systematic model failures and the results obtained for the months of July 87 - June 88 (the data for these months were used in the estimation process) are similar to the results obtained for the other 12 months. Notice that since practically all the QMV coefficients follow a random walk model, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = \hat{\underset{\sim}{\alpha}}^{F}_{t-1}$ so that, for example, the relatively large negative residual mean observed for the month of November 1987 is reflected by a large negative prediction error mean in the month of December 1987.





Figures 7 and 8 show the monthly estimators of the intercept and the floor area coefficient as obtained by ordinary least squares using only the data for the corresponding months, by using the filtered estimators and by using the smoothed estimators. As can be seen the filtered and the smoothed estimators are generally very similar (they are obviously much more apart in the first months not shown in the plots) and they vary only slightly form one month to the other. The OLS estimator on the other hand exhibits a large month to month variation and in the months of July 87 - October 87 the area coefficient estimators came out even negative.



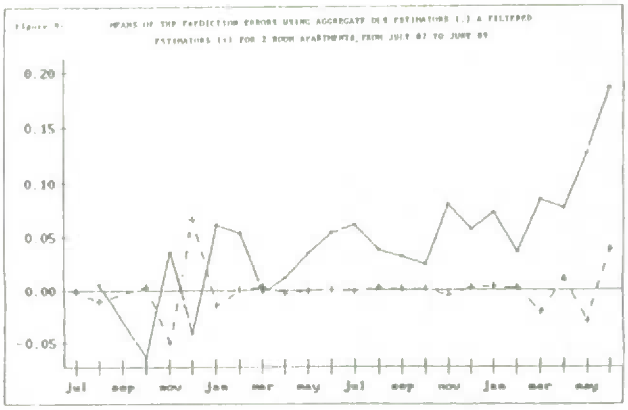


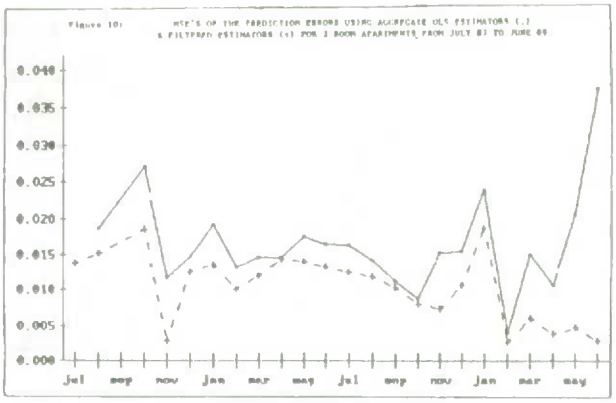
The instability of the OLS estimators is further illustarated in the followiing table 1 where we compare the variances of the OLS and the smoothed estimators for the months of April and May, 1989. As could be expected, the smoothed estimators which use the data of all the months have in all cases much smaller variances.

Table 1: Variances of OLS and Smoothed Estimators of the Regression Coefficients

Month	Estimator	Intercept	Floor Area	Age	District 1	District 2
April 89	OLS	.174	.064	.006	.011	.011
	Smoothed	.068	.025	.0002	.002	.0017
May 89	OLS	.471	.142	.0021	.011	.010
	Smoothed	.093	.033	.0003	.002	.0013

The small month to month variation of the filtered and smoothed estimators could suggest that the regression equations are practically fixed over time. We already mentioned that the variances of the residual terms of the regression coefficients came out highly significant indicating that a model which permits the regression to change over time is more appropriate. In order to further illustrate this point, we compare in figures 9 and 10 the means and MSE's of the prediction errors as obtained by using the filtered estimators (same as in figures 4 and 6) and by using aggregate OLS estimators based on all the data up to and including time t. The plots in these figures are illuminating and they reveal that fixing the regression coefficients over time results in large and increasing prediction biases which translate into increasing prediction MSE's.

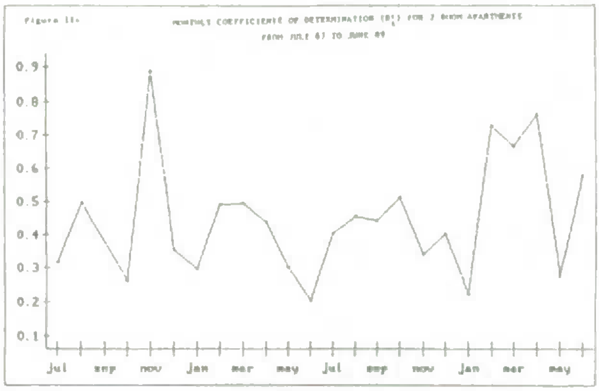


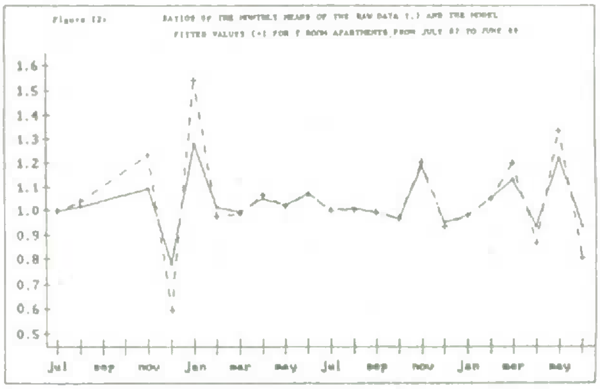


The most important question concerning the goodness of fit of the model is its peformance in estimating the HPI's. In order to partially answer this question we computed two sets of statistics: i) Monthly "coefficients of determination" (R^2) defined as

$$R_t^2 = 1 - \{ \sum_{j=1}^{n_t} [y_{tj} - \exp(\underset{\sim}{x}'_{tj}\, \underset{\sim}{\alpha}_t^F)]^2 \;/\; \sum_{j=1}^{n_t} (Y_{tj} - \bar{Y}_t)^2 \}$$

where $\bar{Y}_t$ is the mean of the sale prices in month t - the results are plotted in Figure 11, and ii) Ratios of the monthly means of the raw data, $R^r_{t|t-1} = \bar{Y}_t / \bar{Y}_{t-1}$ and of the monthly means of the corresponding fitted values $R^f_{t|t-1} = \bar{f}_t \,/\, \bar{f}_{t-1}$ where $\bar{f}_t = \sum_{j=1}^{n_t} \exp\,(\underset{\sim}{x}'_{tj}\, \hat{\underset{\sim}{\alpha}}{}_t^F) \,/\, n_t$. The two sets of ratios are plotted in Figure 12. Notice that all the above statistics have been computed after transforming back from the logarithmic scale.





As can be seen from Figure 11, the R^2 statistics are in most cases above 0.4 which is quite high with this kind of data. Figure 12 reveals a close correspondance between the monthly ratios of the raw data and the ratios of the fitted values. It should be emphasized that these ratios are not estimates of the HPI since they are not necessarily based on prices of homes of similar qualities. However, the fact that the ratios of the fitted values came out so close to the ratios of the original data is very encouraging.

7. CONCLUDING REMARKS

The results of this study indicate that regression relationships within small cells can be estimated efficiently by modelling the variation of the regression coefficients over time. Obviously, further tests are needed to ascertain the suitability of the model. We are already in the process of applying the full model defined by the equations (3.1)-(3.3) incorporating also the robustness modifications suggested in section 5. Comparing the results of the present article with the results obtained for the full model, with and without the modifications, will provide additional insight as to the performance of the model and the effectiveness of the modifications. It is planned also to test the goodness of fit of the model in predicting the sale prices of homes registered after the publication of the index. The registration dates have not been coded in our current working files which is why this test procedure has not been applied so far.

ACKNOWLEDGEMENTS

This study was completed while the first author was on sabbatical leave from the Hebrew University. The authors are grateful to Statistics Canada for supporting this research under its research fellowship program.

REFERENCES

Adelman, I., and Griliches, Z. (1961), "On an Index of Quality Change," *Journal of the American Statistical Association*, 56, 535-548.

Anderson, B.O.D., and Moore, J.B. (1979), "*Optimal Filtering*," Prentice-Hall, Englewood Cliffs, N.J.

Ansley, C.F., and Kohn, R. (1986), "Prediction Mean Squared Error for State Space Models with Estimated Parameters," *Biometrika*, 73, 467-473.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Castles, I. (1987), "The Australian Consumer Price Index Treatment of Home Ownership Costs," Information paper, Catalogue No. 6441.0, Australian Bureau of Statistics, Belconen ACT 2616.

Cooley, T.F., and Prescott, E.C. (1976), "Estimation in the Presence of Stochastic Parameter Variation," *Econometrica*, 44, 167-184.

Court, A.T. (1939), "Hedonic Price Indexes with Automotive Examples," in *The Dynamics of Automobile Demand*, pp. 99-117. New York: General Motors Corporation.

Dielman, T.E. (1983), "Pooled Cross-Sectional and Time Series Data: A Survey of Current Statistical Methodology," *The American Statistician*, 37, 111-122.

Griliches, Z. (1971), "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change," in *Price Index and Quality Change*, Ed. Zvi Griliches, Harvard University Press, pp. 55-87.

Harvey, A.C. (1981), "*Time Series Models*," Philip Allan, Deddington, Oxford.
(1984), "A Unified View of Statistical Forecasting Procedures," *Journal of Forecasting*, 3, 245-275.

Harvey, A.C., and Phillips, G.D.A. (1979), "Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances," *Biometrika*, 66, 49-58.

Hofsten, E.V. (1952), "*Price Indexes and Quality Change*," Bokforlaget Forum AB, Stockholm.

Hsiao, C. (1974), "Statistical Inference for a Model with Both Random Cross-Sectional and Time Effects," *International Economic Review*, 15, 12-30.

Johnson, L.W. (1977), "Stochastic Parameter Regressions: An Annotated Bibliography," *International Statistical Review*, 45, 257-272.

-- (1980), "Stochastic Parameter Regression: An Additional Annotated Bibliography," *International Statistical Review*, 48, 95-102.

LaMotte, L.R., and McWhorter, A. (1977), "Estimation, Testing and Forecasting with Random Coefficient Regression Models," in *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, pp. 814-817.

Maddala, G.S. (1977), "*Econometrics*," McGraw-Hill, Kogakusta.

Meinhold, R.J., and Singpurwalla, N.D. (1983), "Understanding the Kalman Filter," *The American Statistician*, 37, 123-127.

Pfeffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from a Cluster Sample," *Journal of the American Statistical Association*, 76, 681-689.

Pfeffermann, D., and Smith, T.M.F. (1985), "Regression Models for Grouped Populations in Cross-Section Surveys," *International Statistical Review*, 53, 37-59.

Rosenberg, B. (1973a), "The Analysis of Cross-Section of Time Series by Stochastically Convergent Parameter Regression," *Annals of Economic and Social Measurement*, 2, 399,428.

-- (1973b), "A Survey of Stochastic Parameter Regression," *Annals of Economic and Social Measurement*, 2, 381-397.

Schweppe, F. (1965), "Evaluation of Likelihood Functions for Gaussian Signals," *IEEE Transactions on Information Theory*, 11, 61-70.

Stone, R. (1956), "*Quality and Price Indexes in National Accounts*," Organization for European Economic Co-operation, Paris.

Swamy, P.A.V.B., and Mehta, J.S. (1977), "Estimation of Linear Models with Time and Cross-Sectionally Varying Coefficients," *Journal of the American Statistical Association*, 72, 890-898.

Watson, M.W., and Engle, R.F. (1983), "Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models," *Journal of Econometrics*, 23, 385-400.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

ANALYSIS OF SEASONAL ARIMA MODELS FROM SURVEY DATA

D.A. Binder and J.P. Dick[1]

ABSTRACT

A commonly used model for the analysis of time series models is the seasonal ARIMA model. However, the survey errors of the input data are usually ignored in the analysis. We show, through the use of state-space models with partially improper initial conditions, how to estimate the unknown parameters of this model using maximum likelihood methods. As well, the survey estimates can be smoothed using an empirical Bayes framework. We apply these techniques to an unemployment series from the Labour Force Survey.

1. INTRODUCTION

It is common practice to analyze data from surveys where similar data items are collected on repeated occasions, using time series analysis methods. Most standard methods for these analyses assume the data are either observed without error or have independent measurement errors. However, in the analysis of repeated survey data, when there are overlapping sampling units between occasions, the survey errors can be correlated over time.

A commonly used model in the analysis of time series is the seasonal integrated autoregressive-moving average (ARIMA) regression model, which we discuss in this paper. We show how to incorporate the (possibly correlated) survey errors into the analysis. In particular, we consider the case where the survey (design) error can be assumed to be an ARMA process up to a multiplicative constant.

When such a model for the behaviour of the population characteristics is assumed, the minimum mean squared error, or, equivalently, the Bayes linear estimator for the characteristic at a point in time can be derived. This estimator incorporates the model structure which the classical estimators, such as the minimum variance linear unbiased estimators, ignore. When the model parameters are estimated from the survey data, the estimators are empirical Bayes.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Jones (1980) and others considered the implications of certain stochastic models for the population means over time. In Binder and Dick (1989), these results were generalized using state space models and Kalman filters. In this paper, we extend the framework to include the model where differencing of the original series of the population means yields an ARMA model. We use the modified Kalman filter approach given by Kohn and Ansley (1986). To estimate the unknown parameters, we maximize the marginal likelihood function using the method of scoring. This approach can also handle missing data routinely. We also show how the survey estimates can be smoothed to incorporate the model features using empirical Bayes methods. Confidence intervals for these smoothed values are also given, using the method described by Ansley and Kohn (1986).

An example of this model is described in Section 5 using unemployment data from the Canadian Labour Force Survey. This example shows the implications on the estimates of the model parameters when the survey errors are taken into account. We also derive a smoothed estimate of the underlying process under the model assumptions.

2. THE MODEL

Suppose we have a series of point estimates from a repeated survey of a population characteristic, given by $y_1, y_2, \ldots, y_T$. We assume that y_t can be decomposed into three components, so that

$$y_t = x'_t\gamma + \theta_t + e_t, \tag{2.1}$$

where $x'_t\gamma$ is a deterministic regression term, θ_t is a population parameter following a time series model, and e_t is the survey error, assumed to have zero expectation.

We first describe an integrated seasonal autoregressive-moving average model for $\{\theta_t\}$. We let B be the backshift operator; $\nabla = 1-B$ and $\nabla_s = 1-B^s$, where s is the seasonal period. We define the following polynomial functions:

$$\lambda(A) = 1 - \lambda_1 A - \lambda_2 A^2 - \ldots - \lambda_p A^p,$$

[1] D.A. Binder, Business Survey Methods Division and J.P. Dick, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

$$\alpha(A) = 1 - \alpha_1 A - \alpha_2 A^2 - \ldots - \alpha_p A^p,$$

$$\nu(A) = 1 - \nu_1 A - \nu_2 A^2 - \ldots - \nu_Q A^Q,$$

and

$$\beta(A) = 1 - \beta_1 A - \beta_2 A^2 - \ldots - \beta_q A^q.$$

The seasonal ARIMA $(p,d,q)(P,D,Q)_s$ model for $\{\theta_t\}$ is given by

$$\lambda(B^s)\alpha(B)\nabla^d\nabla_s^D\theta_t = \nu(B^s)\beta(B)\varepsilon_t, \tag{2.2}$$

where the ε_t's are independent $N(0,\sigma^2)$. We define $a(B) = \lambda(B^s)\alpha(B)$, a $(p+sP)$-degree polynomial; $\Delta(B) = \nabla^d\nabla_s^D$, a $(d+sD)$-degree polynomial; $b(B) = \nu(B^s)\beta(B)$, a $(q+sQ)$-degree polynomial; $A(B) = a(B)\Delta(B)$, a $(p+d+sP+sD)$-degree polynomial; $u_t = \Delta(B)\theta_t$, an ARMA$(p+sP,q+sQ)$ process. Therefore, alternative representations of (2.2) are

$$a(B)\Delta(B)\theta_t = b(B)\varepsilon_t, \tag{2.3}$$

$$A(B)\theta_t = b(B)\varepsilon_t, \tag{2.4}$$

and

$$a(B)u_t = b(B)\varepsilon_t, \tag{2.5}$$

We now consider the survey errors $\{e_t\}$ of expression (2.1). It will be assumed that the sample sizes of the repeated survey are sufficiently large that the errors for the survey estimates can be approximated by a multivariate normal distribution. In the simplest case, where the surveys are non-overlapping and the sampling fractions are small, the e_t's can be assumed to be independent. In a rotating panel survey, the survey errors are usually correlated. In this case, since the correlations between survey occasions are zero after panels have been rotated out, a pure moving average process can be used to describe the survey error process.

Alternatively, if a random sample of units are replaced on each survey occasion, a pure autoregressive process may best describe the process. More complicated models are also possible. For example, in a two-stage design, some of the first stage units may be replaced randomly on each occasion and the second stage units may have a rotating panel design. This might be represented by an autoregressive-moving average process.

In this paper, we assume that the survey error process is given by

$$e_t = k_t\omega_t, \tag{2.6}$$

where $\{\omega_t\}$ is an ARMA (m,n) process, given by

$$\phi(B)\omega_t = \psi(B)\eta_t \tag{2.7}$$

and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_m B^m,$$

and

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \ldots - \psi_n B^n,$$

The η_t's are independent $N(0,\tau^2)$. The factor k_t has been included in (2.6) to allow for non-homogeneous variances, even when the autocorrelation function is homogeneous in time.

In the model just described we assume that τ^2, the k_t's and the coefficients of $\phi(B)$ and of $\psi(B)$ can be estimated directly from the survey data, using design-based methods. However, in general, the other parameters are unknown. This includes γ, σ^2, and the coefficients of $\lambda(A)$, $\alpha(A)$, $\nu(A)$ and of $\beta(A)$. The x_t's is the regression term are assumed known.

3. STATE SPACE FORMULATION OF THE MODEL

3.1 General Formulation

The model described in Section 2 can be formulated as a state space model with partially improper priors. This has a number of advantages. It permits, through use of a modified Kalman filter, calculation of a marginal likelihood function, which can be maximized to estimate unknown parameters. It also accommodates smoothing of the original survey estimates, by removing the estimates of survey error from the data.

In the state space model, two processes occur simultaneously. The first process, the observation system, details how the observations depend on the current state of the process parameters. The second process, the transition system, details how the parameters evolve over time.

For the state space models we consider here, the observation equation is written as

$$y_t = h_t' z_t \tag{3.1a}$$

and the transition equation is

$$z_t = F z_{t-1} + G \xi_t, \tag{3.1b}$$

where z_t is an $(r \times 1)$ state vector and h_t is a fixed $(r \times 1)$ vector. In the transition equation, F is a fixed $(r \times r)$ transition matrix, G is a fixed $(r \times m)$ matrix and the ξ_t's are independent normal vectors with mean zero and covariance U.

The final requirement to complete the specification of the state space process is the initial conditions for z_0. In this paper, we shall use the improper prior formulation given in Kohn and Ansley (1986). In general, we assume that z_0 has a partially diffuse r-variate normal distribution with mean $m(0|0) = 0$ and covariance matrix $V(0|0)$, where

$$V(0|0) = \kappa V_1(0|0) + V_0(0|0) \tag{3.2}$$

for large κ.

We denote the conditional mean of z_t given the observations up to and including time t' by $m(t|t')$, and the conditional variance by $V(t|t')$, where

$$V(t|t') = \kappa V_1(t|t') + V_0(t|t'). \tag{3.3}$$

Recursive formulae for the cases where $t=t'$ and $t=t'+1$ are given in Kohn and Ansley (1986). They refer to this as the modified Kalman filter.

Since the model for $\{y_t\}$ given by (2.1) contains survey errors $\{e_t\}$ an estimate of the components without survey error, given by

$$y_t \text{ (smoothed)} = x_t' \gamma + \theta_t \tag{3.4}$$

is often of interest. When the right hand side of (3.4) can be expressed as $g_t' z_t$, for some g_t', then it is possible to obtain the conditional mean and variance of the linear combination $g_t' z_t$ given all the data, using the modified Kalman filter. To do this, the recursions are applied up to time t to obtain $m(t|t)$ and $V(t|t)$. Then the state vector z_t is augmented by the state $z_{t,r+1} = g_t' z_t$, and $m(t|t)$ and $V(t|t)$ are also appropriately augmented. The matrix F in (3.1b) is modified to add the equation $z_{t+1,\ r+1} = z_{t,r+1}$. After these modifications, the modified Kalman filter can be used as before sc that the last component of $m(T|T)$ gives the conditional expectation of $g_t' z_t$, given all the data, $y_1, y_2, \ldots y_T$. As well, the last diagonal component of $V(t|t)$ gives the conditional variance. This procedure can be generalized to include any number of smoothed estimates and their conditional covariances.

3.2 Model for θ

Harvey and Phillips (1979) described a method to put the ARIMA model (2.4) into the state space form given by (3.1). The dimension of z_t is $r = \max(p+d+sP+sD,\ q+sQ)$. By augmenting $A = (A_1, \ldots, A_{p+d+sP+sD})$ or

$\mathbf{b} = (b_1, \ldots, b_{q+sQ})$ with zeroes to have dimension r, the ARIMA model may be written in the form given by (3.1), where $\mathbf{h}_t' = (1, 0, \ldots, 0)$, $\mathbf{G}_t' = (1, -b_1, \ldots, -b_{r-1})$ and

$$F = \left|\begin{array}{c|c} A_1 & \\ \vdots & I_{r-1} \\ A_{r-1} & \\ \hline A_r & 0' \end{array}\right|,$$

where I_{r-1} is the $(r-1) \times (r-1)$ identity matrix and 0' is a row vector of zeroes.

In this formulation, the state vector $\mathbf{z}_t = (z_{1t}, \ldots, z_{rt})'$ is defined as

$$\begin{aligned} z_{it} &= A_i\theta_{t-1} + A_{i+1}\theta_{t-2} + \ldots + A_r\theta_{t-(r-i+1)} \\ &\quad - b_{i-1}\epsilon_t - b_i\epsilon_{t-1} - \ldots - b_{r-1}\epsilon_{t-(r-i)}, \end{aligned} \tag{3.5}$$

for $i = 2, 3, \ldots, r$ and $z_{1t} = \theta_t$.

To complete the specification for $\{\theta_t\}$, initial conditions for z_0 are required. These are given in Ansley and Kohn (1985), a summary of which is provided here.

From expression (2.5), $\{u_t\}$ is an ARMA process. We define

$$\theta\_ = (\theta_0, \theta_{-1}, \ldots, \theta_{-S})',$$

where $S = \max(0, p+sP+d+sD-1)$. We let

$$u\_ = (u_0, u_{-1}, \ldots, u_{-R})',$$

where $R = \max(0, p+sP-1)$. Finally, we let

$$w\_ = (\theta_{-R-1}, \theta_{-R-2}, \ldots, \theta_{-S})',$$

when $S > R$.

Now, $\mathbf{u}\_$ is assumed to be a stationary ARMA process, so that its covariance matrix can be derived from expression (2.5). It is assumed that $\mathbf{w}\_$ is $N(0, \kappa I)$ and is independent of $\mathbf{u}\_$. Since $(\mathbf{u}\_', \mathbf{w}\_')'$ is a linear combination of $\theta\_$, the covariance matrix for $\boldsymbol{\theta}$ can be derived. Using the form of expression (3.5) for z_0, the initial covariance matrix can be computed. Note that when both d and D are zero, so that no differencing takes place in the model, then $\mathbf{w}\_$ is the null vector and we have $\mathbf{u}\_ = \boldsymbol{\theta}\_$.

3.3 Model for the Observed Data

In Section 2 we assumed that $e_t = k_t\ \omega_t$, where ω_t is an ARMA(m,n) model. Therefore, from the discussion in Section 3.3, it is clear that e_t can be represented in state space form, with $\mathbf{h}_t = (k_t, 0, \ldots, 0)'$, and $e_t = \mathbf{h}_t'z_t$.

The regression component can be similarly represented. We let $z_0 = \gamma$, the regression coefficients, assumed to have mean zero and covariance κI. The transition equation is simply $z_{t+1} = z_t$.

Since we can represent each of the components of y_t in expression (2.1) by a state space model, it straightforward to combine the individual models into an overall model, by extending the state vector to include the state vectors from the individual components. The observation equation is then the sum of the three individual components.

4. ESTIMATION OF THE STATE SPACE MODEL

4.1 Estimation of the Parameters

The unknown parameters of this model are σ^2, and the coefficients of $\lambda(A)$, $\alpha(A)$, $\nu(A)$ and $\beta(A)$. We performed the iterations on $\log(\sigma^2)$, rather than σ^2, to avoid problems with negative values. Note that the regression coefficients, γ, are included as parameters of the state vector. The model for the vector of observations $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ given in Section 3 is equivalent to

$$y = M\eta + \zeta, \tag{4.1}$$

where η is j-variate $N(0, \kappa I)$, ζ is T-variate $N(0, \mathbf{W})$, and $\mathbf{M}$ is a $T \times j$ matrix.

Kohn and Ansley (1986) recommended maximizing the limit of $\kappa^{j/2}$ times the likelihood function for the data, as κ tends to infinity. It can be shown that the limit of the likelihood function is equivalent to the marginal likelihood function of $\mathbf{y} - \mathbf{M}\hat{\eta}$, where $\hat{\eta}$ is the maximum likelihood estimate of η when $\mathbf{M}$ and $\mathbf{W}$ are known. Tunnicliffe-Wilson (1989) has shown that the Jacobian of transformation from the data $\mathbf{y}$ to $(\hat{\eta}, \; \mathbf{y} - \mathbf{M}\hat{\eta})$ does not depend on the model parameters of $\mathbf{W}$ whenever $\mathbf{M}$ is known. As well, the derivative of the transformation from $\mathbf{y}$ to $\hat{\eta}$ is $\mathbf{M}$. Ansley and Kohn (1985) has shown that M does not depend on the unknown parameters. By using the modified Kalman filter, the computations for the marginal likelihood function are straightforward.

The procedure we employed computes both the marginal likelihood function and its first derivatives with respect to the unknown parameters. This involves taking first derivatives of the initial conditions and of $\mathbf{m}(t|t')$ and the components of $\mathbf{V}(t|t')$ for $t=t'$ and $t=t'+1$. All the computations were done using PROC IML in SAS.

The likelihood function was maximized using a modification of the method of scoring. This modification allowed for varying step sizes. On each iteration, the likelihood function was computed at the previous step size, as well as at this step size multiplied and divided by a predetermined constant. (We used 1.1 as the factor.) The next step size was that which maximized the likelihood function among the three points. Each time a check was made to determine whether the parameters were in range. This was done by checking for positive semi-definiteness of the initial covariance matrix of the state vector. If it was out of range, the step size was divided again by the constant and the procedure repeated.

To obtain the estimated variance matrix for the estimated parameters, the inverse of the Fisher information was used. This is readily computed since the first derivatives of the likelihood function are available.

4.2 Estimation of the Smoothed Values

Smoothed values for the estimates can be obtained by zeroing out that component of the state vector which corresponds to the survey error. However, this still leaves open the question of how to estimate its variance. To derive the standard error of the smoothed estimate it is necessary to account for the fact that the unknown parameters have been estimated from the data, particularly when the data series is short; see Jones (1979).

To obtain the variance of $\mathbf{g}'\mathbf{z}_t$, it is sufficient to derive the variance $\mathbf{z}_T - \hat{\mathbf{m}}(T|T)$, where $\hat{\mathbf{m}}(T|T)$ is the estimate of $\mathbf{m}(T|T)$ at the estimated parameter values. This is because the state vector has been augmented to include $\mathbf{g}'\mathbf{z}_t$. Now,

$$\begin{aligned} z_T - \hat{\mathbf{m}}(T|T) &= [z_T - \mathbf{m}(T|T)] \\ &\quad + [\mathbf{m}(T|T) - \hat{\mathbf{m}}(T|T)]. \end{aligned} \tag{4.2}$$

The first component of the right hand side of (4.2) has conditional variance $\mathbf{V}(T|T) = \mathbf{V}_0(T|T)$, assuming that $\mathbf{V}_1(T|T) = 0$. The second component of (4.2) represents a bias term and is independent of the first term, since it depends only on the data y. By taking a Taylor series expansion of the second term around the true parameter values and ignoring higher terms, we have the second component of (4.2) is

$$\mathbf{m}(T|T) - \hat{\mathbf{m}}(T'T) = \left[\frac{-\partial\hat{\mathbf{m}}(T|T)}{\partial\hat{\phi}}\right]' (\hat{\phi} - \phi), \tag{4.3}$$

where ϕ is the vector of unknown parameters and $\hat{\phi}$ is its estimate. Therefore, the variance of (4.2) is approximately

$$\mathrm{Var}[z_T - \hat{m}(T|T)] = V_0(T|T) + [\frac{\partial \hat{m}(T|T)}{\partial \phi}]' V_\phi [\frac{\partial \hat{m}(T|T)}{\partial \phi}] \ , \qquad (4.4)$$

where V_ϕ is the covariance matrix for the unknown parameters. Expression (4.4) is estimated by using the estimated parameter values. This is the same approach as that given by Ansley and Kohn (1986).

5. LABOUR FORCE SURVEY DATA

To demonstrate this procedure, we took data from the Canadian Labour Force Survey (LFS). The LFS is a monthly rotating panel survey. Each panel, which contains one-sixth of the selected households, remains in the sample for six consecutive months: the sample design is a stratified multi-stage design. The primary sampling units are rotated out after approximately two years.

The data were the estimated monthly number of unemployed from January 1977 to December 1986 in Nova Scotia and the subprovincial area within Nova Scotia corresponding to Cape Breton Island. This province was chosen because the sampling errors were moderate compared to the larger provinces and because subprovincial data were available. The logarithm of the Nova Scotia data is displayed on Graph 1a while the logarithm of the Cape Breton Island data is shown on Graph 2a. The models were fitted to this transformed series.

Lee (1987) estimated the autocorrelations for Nova Scotia survey error process up to a lag of eleven. Using these autocorrelations, we used the method of moments to estimate the coefficients of τ^2, $\phi(B)$ and $\psi(B)$ given in (2.7). A good fit was found using an ARMA(3,6) model. The estimated parameters were $\phi_1 = 0.2575$, $\phi_2 = -0.358$, $\phi_3 = -0.6041$, $\psi_1 = -0.1847$, $\psi_2 = -0.5873$, $\psi_3 = 0.3496$, $\psi_4 = 0.0647$, $\psi_5 = 0.0982$, $\psi_6 = 0.0347$, and $\tau^2 = 0.7246$. The k_t's of (2.6) were the estimated standard errors of the estimates, derived by taking a Taylor series approximation for the logarithms.[2]

A series of models were fitted to the data where no sampling error was assumed; that is, all the k_t's were taken as zero. These models were then refitted using the assumed structure for the survey error. We compared the estimated parameter values. As well in the case where the survey error structure is assumed to be non-zero, we computed smoothed values for the survey estimates and compared their standard errors with the standard errors of the original series.

Initially the model selected for the Nova Scotia series incorporating the survey error, was a seasonal ARIMA $(1,1,0)(0,0,1)_{12}$) with a deterministic regression term to account for the seasonality. The 12 regression variables included a linear term and a dummy variable for each of the first 11 months. The dummy variable for a reference month took the value 1 for the reference month, -1 for December and 0 for the other months. Note that an intercept term is not estimable because the first differences of the data are fitted. The estimated parameters for this model were highly unstable so it was decided to drop the seasonal moving average component from the model. This left as the model an ARIMA (1,1,0) with a deterministic regression term. The same model was used to for the Nova Scotia data ignoring the sampling error and for the Cape Breton Island data.

The parameter estimates for both Nova Scotia and Cape Breton Island are displayed in Table 1. We display the estimates which do not take into account the survey error component in the "Without Sampling Error" columns. The estimates from both models for Cape Breton Island, especially for the regression estimates, are very similar. Note that the AR component also has a similar estimates and that the 'With Sample Error' model has reduced the variance substantialy. The column headed by 'T-value' displays the test statistics for assuming a true value of zero for the parameter. Note that the significance level for the regression estimates is fairly close in every case. However, the model 'Without Sampling Error' indicates a strong significance level ($t = -2.85$) for the AR(1) component while the model incorporating the survey error process shows no need to include the AR component in the model ($t = -0.68$). This result leads to accepting a regression model for the unemployment series in Cape Breton Island for the model with the survey process incorporated. If the survey error is ignored, then the model would include a term relating the previous month's estimate to the current month's.

The results for Nova Scotia have some similarities to the Cape Breton Island results. The regression estimates for both the 'With error' and 'Without error' models are fairly close. Note that the significance level for the regression estimates in the 'With error' model are much smaller than in the 'Without error' model. The variance reduction for the 'With error' model relative to the 'Without error' model is far larger than the variance reduction between the same two models for the Cape Breton Island data. However, the most interesting result is in the AR component. Both models show that the AR component is significant for each model. The estimates, however, are entirely different. The 'Without error' model gives as an estimate of $\alpha = -0.296$. The 'With error' model estimates $\alpha = 0.862$. Clearly, the intrepretations that would be made are different.

Intuitively, after removing the trend and the monthly effects, it would be expected that the previous month's estimate would have a positive correlation with the current month. This is exactly what happens in the 'With error' model. It would seem that the negative AR component estimated for the model 'Without error' is picking up some of survey error process; thus leading to a misleading intrepretation to the data.

Graph 1a shows the smoothed estimates calculated from the model incorporating survey errors superimposed on the original data points for Nova Scotia. Graph 2a shows similar smoothed estimates for Cape Breton Island. The observed values minus the smoothed estimates for the Nova Scotia series are displayed in Graph 1b. From this graph it can be seen that the recession of 1981 is having a large impact. Prior to 1981 the smoothed estimates tend to be higher than the original values while after 1981 the smoothed estimates tend to be lower than the original values. The observed minus the smoothed estimator from Cape Breton Island are displayed in Graph 2b. These appear to form a more random arrangement than the Nova Scotia results probably due to the larger sampling errors associated with the Cape Breton Island data.

In summary, when the sampling error component is incorporated, the best model can differ from the case when sampling is ignored or it can provide an entirely different interpretation to the model. The data from Cape Breton Island displays a situation when the survey error is accounted for, where a regression model will satisfactorily explain the data while the model ignoring the survey process requires the inclusion an AR component. On the other hand, the Nova Scotia data required an AR component for both models, but gave entirely different interpretations for these components. In the future more work needs to be developed on evaluating the competing models. In particular, since the one-step ahead prediction errors can be combined with the estimates to form a independent normal process, these predictions can be evaluated using standard residual analysis procedures. Future work will detail the results of incorporating this analysis.

REFERENCES

Ansley, C.F., and Kohn, R. 1985. A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. *Journal of Statistical Computation and Simulation*. 21: 135-169.

Ansley, C.F., and Kohn, R. 1986. Prediction mean squared error for state space models with estimated parameters. *Biometrika*. 73: 467-473.

Binder, D.A., and Dick, J.P. 1989. Modelling and estimation for repeated surveys. *Survey Methodology*. 14: 29-46.

Blight, B.J.N., and Scott, A.J. 1973. A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B. 35: 61-68.

Harvey, A.C., and Phillips, G.D.A. 1979. Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*. 66: 49-58.

Jones, R.G. 1979. The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*. 21: 45-56.

Jones, R.G. 1980. Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, Series B. 42: 221-226.

Kohn, R., and Ansley, C.F. 1986. Estimation, prediction and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*. 81: 751-761.

Lee, H. 1987. Estimation of panel correlations for the Canadian Labour Force Survey. *Technical Report* SSMD-89-023E. Statistics Canada.

Scott, A.J., and Smith, T.M.F. 1974. Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*. 69: 674-678.

Scott, A.J., Smith, T.M.F., and Jones, R.G. 1977. The application of time series methods to the analysis of repeated surveys. *International Statistics Review*. 45: 13-28.

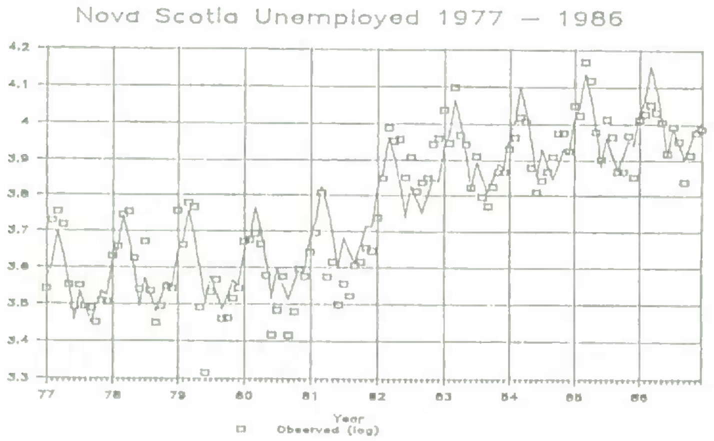
Tunnicliffe-Wilson, G. 1989. On the Use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Series B. 51: 15-27.

Table 1

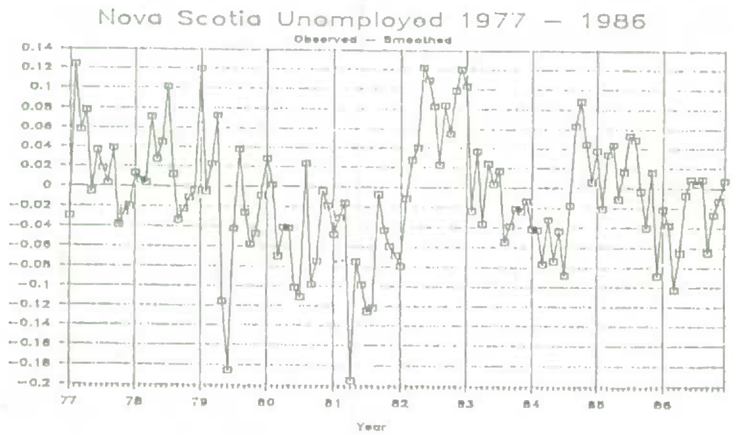
Paramter Estimates - Unemployment Series 1977 -1985

	Nova Scotia				Cape Breton Island			
	Without Sampling Error		With Sampling Error		Without Sampling Error		With Sampling Error	
Parameter	Estimate	T-value	Estimate	T-value	Estimate	T-value	Estimate	T-value
Alpha (1)	-0.296	-3.23	0.862	2.08	-0.260	-2.85	-0.231	-0.68
Sigma	0.0597	-	0.0032	-	0.1049	-	0.0520	-
Trend	0.00427	1.01	0.00420	1.89	0.00607	0.79	0.00598	1.50
January	0.064	3.60	0.048	1.93	-0.007	-0.23	-0.003	-0.10
February	0.083	4.80	0.078	3.30	0.027	0.89	0.028	0.97
March	0.166	10.20	0.165	6.40	0.171	5.76	0.164	5.76
April	0.106	6.60	0.104	4.10	0.099	3.33	0.089	3.19
May	0.009	0.60	0.016	0.70	-0.008	-0.28	-0.007	-0.24
June	-0.101	-6.00	-0.088	-3.30	-0.029	-0.96	-0.033	-1.17
July	-0.016	-1.20	-0.014	-0.63	0.082	2.77	0.081	3.13
August	-0.058	-3.60	-0.062	-2.37	-0.011	-0.37	-0.009	-0.30
September	-0.106	-6.60	-0.105	-3.96	-0.104	-3.51	-0.098	-3.18
October	-0.081	-4.80	-0.071	-3.08	-0.084	-2.83	-0.069	-2.44
November	-0.026	-1.80	-0.029	-1.08	-0.063	-2.10	-0.074	-2.46

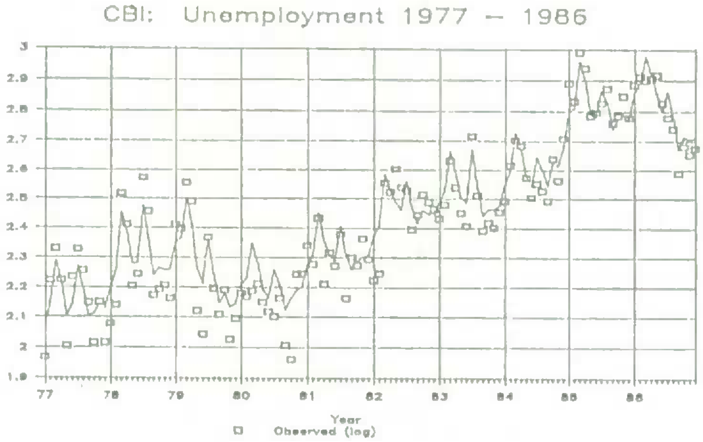
Graph 1a



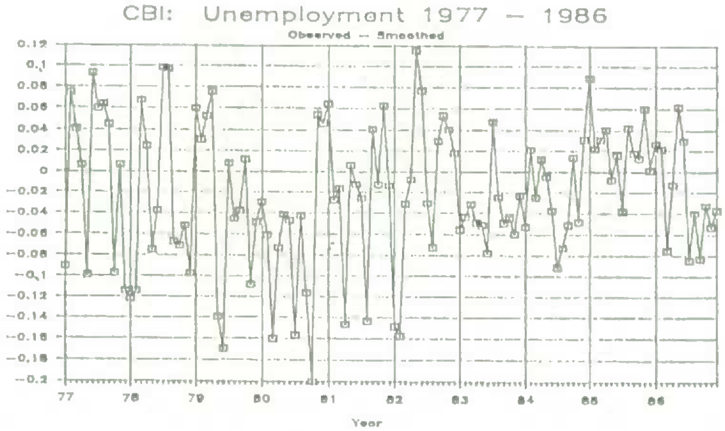
Graph 1b



Graph 2a



Graph 2b



Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

SMALL AREA ESTIMATION USING MODELS THAT COMBINE TIME SERIES AND CROSS-SECTIONAL DATA

G.H. Choudhry[1] and J.N.K. Rao[2]

ABSTRACT

Cross-sectional and time series models with random effects and autocorrelated errors are developed. Using these models, "best linear unbiased" estimators for small areas at each time point are obtained. The efficiencies of several small area estimators are evaluated, using monthly survey estimates of unemployment for census divisions (small areas) from the Canadian Labour Force Survey in conjunction with monthly administrative counts from the Unemployment Insurance System and monthly survey estimates of population in labour force as auxiliary variables.

1. INTRODUCTION

The demand for reliable small area statistics has steadily increased in recent years due to their use in formulating policies and programs, in allocation of government funds, and in regional programs. Statistics Canada responded to user needs by undertaking a program of small areas development. Brackstone (1986) discussed the issues arising in the development and provision of small area data.

Direct small area estimators from survey data are likely to yield unacceptably large standard errors due to small sample sizes. Alternative estimators that "borrow strength" from related small areas are therefore needed to improve efficiency. Such estimators use models, either implicitly or explicitly, that link the small areas through supplementary data such as recent census counts and administrative records.

Most of the research on small area estimation has focused on cross-sectional data at a given point in time. Rao (1986) has given an account of this research. Estimators proposed in the literature include (a) synthetic estimators (Gonzalez, 1973; Ericksen, 1974), structure preserving estimators (SPREE), Purcell and Kish (1980); (b) sample size dependent estimators (Drew et al. 1982; Särndal and Hidiroglou, 1989); (c) empirical Bayes estimators (Fay and Herriot, 1979) and empirical best linear unbiased predictors (EBLUP), Battese et al. (1988) and Prasad and Rao (1990). The EBLUP is obtained from the best linear unbiased predictor (BLUP) by replacing the unknown variance parameters with their estimates, similar to the empirical Bayes estimator obtained from the Bayes estimator.

The main purpose of this paper is to develop cross-sectional and time series models with random effects and autocorrelated errors, and to obtain EBLUP's for small areas at each point in time using these models. Section 2 reviews the work on regression synthetic estimators and empirical Bayes estimators obtained from cross-sectional data at a given point in time. Cross-sectional and time series models are considered in Section 3, and an extension of the Fay-Herriot (1979) model is proposed. The EBLUP is obtained in Section 4. The efficiencies of EBLUP, relative to two synthetic estimators and a direct survey estimator are evaluated in Section 5, using monthly survey estimates of unemployment for census divisions (small areas) from the Canadian Labour Force Survey in conjunction with monthly administrative counts from the Unemployment Insurance (UI) system and monthly survey estimates of population in labour force as auxiliary variables.

2. CROSS-SECTIONAL ESTIMATORS

2.1 Regression Synthetic Estimators

Let y_i be the direct survey estimator of i-th small area mean θ_i at a given point in time. For simplicity, we assume that a single concomitant variable x_i related to θ_i is available; extension to two or more concomitant variables is straightforward. We also assume that y_i is unbiased for θ_i, i.e., $y_i = \theta_i + e_i$ where the e_i's are the sampling errors with $E(e_i) = 0$.

We assume the following linear regression model on the θ_i's that links the small areas through the concomitant data x_i:

$$\theta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, I, \tag{2.1}$$

[1] G.H. Choudhry, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

[2] J.N.K. Rao, Department of Mathematics & Statistics, Carleton University, Ottawa, Ontario K1S 5B6

where β_0 and β_1 are the regression coefficients. A regression synthetic estimator of θ_i is then given by

$$\tilde{\theta}_{i(\mathrm{reg})} = \tilde{\beta}_0 + \tilde{\beta}_1 x_i, \tag{2.2}$$

where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the ordinary least squares estimators of β_0 and β_1 obtained from the combined model $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \ldots, I$. Alternatively, we can use the generalized (weighted) least squares estimators of β_0 and β_1 if the estimated covariance matrix of survey estimators y_i is available.

Synthetic estimator (2.2) could lead to large biases since it does not give a weight to the direct survey estimator y_i. On the other hand, the empirical Bayes estimator or the EBLUP gives proper weights to the survey estimator and the synthetic estimator, and as a result leads to smaller biases relative to the synthetic estimator.

2.2 Empirical Bayes Estimator or EBLUP

Fay and Herriot (1979) introduced uncertainty into the model (2.1) as follows:

$$\theta_i = \beta_0 + \beta_1 x_i + v_i, \tag{2.3}$$

where the v_i's are independent normal variables with mean 0 and unknown variance σ_v^2. For sampling errors, they assumed that the e_i's are independent normal variables with $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma_i^2$, where σ_i^2 is known. The combined model is given by

$$y_i = \beta_0 + \beta_1 x_i + v_i + e_i. \tag{2.4}$$

The empirical Bayes estimator of θ_i is given as a weighted sum of the direct survey estimator y_i and the regression synthetic estimator $\hat{\theta}_{i(\mathrm{reg})} = \hat{\beta}_0 + \hat{\beta}_1 x_1$:

$$t_i(\hat{\sigma}_v^2, \mathbf{y}) = w_i y_i + (1 - w_i)\hat{\theta}_{i(\mathrm{reg})}, \tag{2.5}$$

where $w_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \sigma_i^2)$, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the weighted least squares estimators under the combined model, and $\hat{\sigma}_v^2$ is an estimator of σ_v^2. A simple moment estimator of σ_v^2 or a more complicated estimator, such as the maximum likelihood estimator of σ_v^2, may be used. Fay and Herriot (1979) used (2.5) to estimate per capita income for small areas (i.e., population less than 1000) from the 1970 U.S. Census of Population and Housing, and presented evidence that (2.5) leads to smaller average error than either the direct survey estimator or the synthetic estimator using the county average.

Prasad and Rao (1990) obtained an accurate estimator of the mean squared error of EBLUP (2.5) by taking account of the uncertainty in the estimator of σ_v^2.

3. CROSS-SECTIONAL AND TIME SERIES MODELS

The methods of Section 2 use only cross-sectional data at a given point in time, and as a result do not exploit information in data at other time points. Scott et al (1977), Jones (1980), Tiller (1989) and others used time series modelling of aggregates (e.g., overall means) from repeated survey data, and obtained improved estimators of aggregates at different time points. However, very little work has been reported on extending the Fay-Herriot approach for small area estimation to time series of cross-sectional survey estimates of small areas in conjunction with census data and time varying supplementary data such as administrative records.

Cronkhite(1986) developed regression synthetic estimators using pooled cross-sectional time series data, and applied them to estimate substate area employment and unemployment, using the Current Population Survey (CPS) monthly survey estimates as dependent variable and counts from the UI system and census variables as independent variables. The motivation for our research was to obtain reliable monthly estimates of unemployment for census divisions, using Labour Force Survey estimates of unemployment and labour force participation rates, and administrative counts from the UI system. Three-year average unemployment rates for census divisions are used in conjunction with other variables to produce an index which in turn is used to allocate funds for industrial incentive.

Extensive econometric literature exists on modelling and estimating relationships that combine time series and cross-sectional data (for example, see Judge et al, 1980, Chapter 13), but sampling errors are seldom taken into account. We now consider some of these models. For simplicity, we again consider only one concomitant variable. Let θ_{it}, y_{it} and x_{it} respectively be the population mean, the direct survey estimate and the concomitant variable associate with the i-th small area at time t ($i = 1, \ldots, I$; $t = 1, \ldots, T$). We have

$$y_{it} = \theta_{it} + e_{it}, \quad i = 1, \ldots I; \quad t = 1, \ldots, T, \tag{3.1}$$

and, following Fay and Herriot (1979), we assume that the covariance matrix of sampling errors e_{it} is block diagonal with known blocks Σ_i, where Σ_i is a $T \times T$ matrix, and $E(e_{it}) = 0$. Recent research has focused on modelling sampling errors of aggregates. For example, Binder and Dick (1989) and Tiller (1989) proposed autoregressive moving average (ARMA) models.

The models on θ_{it}, proposed in the econometric literature, include the following:

$$\text{(I)} \qquad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \epsilon_{it}, \tag{3.2}$$

where the v_i's are fixed small area effects and the ϵ_{it}'s are independent normal variables with mean 0 and variance σ^2, abbreviated $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$.

$$\text{(II)} \qquad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \epsilon_{it}, \tag{3.3}$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$, $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$ and $\{v_i\}$ and $\{\epsilon_{it}\}$ are independent. Here the v_i's are random small area effects.

$$\text{(III)} \qquad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_t + \epsilon_{it}, \tag{3.4}$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$, $u_t \sim_{\text{ind}} N(0, \sigma_u^2)$, $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$ and $\{v_i\}$, $\{u_t\}$, $\{\epsilon_{it}\}$ are independent. Here v_i's and u_t's are random small area effects and random time effects respectively

$$\text{(IV)} \qquad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_{it}, \tag{3.5}$$

and

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}, \quad |\rho| < 1$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$, $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$ and $\{v_i\}$, $\{\epsilon_{it}\}$ are independent. Here the v_i's are random small area effects and $\{u_{it}\}$ follow an AR(1) process. The model (3.5) may be rewritten as a distributed lag model:

$$\theta_{it} = \rho\theta_{i,t-1} + (1-\rho)\beta_0 + \beta_1 x_{it} - \beta_1 \rho x_{i,t-1} + (1-\rho)v_i + \epsilon_{it}. \tag{3.6}$$

Model IV appears to be the most realistic among the four models since the alternative form (3.6) relates the current population mean, θ_{it}, to the previous period population mean, $\theta_{i,t-1}$, and to the values of the auxiliary variable for the current and previous periods, x_{it} and $x_{i,t-1}$ respectively. The form (3.5) of model IV reflects the dependence of θ_{it} over time for each area i. Henceforth, we adopt model IV in the form (3.5).

The combined model using (3.1) and (3.5) is given by

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + v_i + (e_{it} + u_{it}), \\ u_{it} &= \rho u_{i,t-1} + \epsilon_{it}, \ |\rho| < 1, \end{aligned} \tag{3.7}$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$ $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$, and the e_{it}'s have mean zero and known block diagonal covariance matrix $\Sigma = \text{diag}(\Sigma_1, \ldots, \Sigma_I)$.

Unfortunately, the sampling covariance matrix Σ from the Canadian Labour Force Survey is currently not available, so we treated the composite error $w_{it} = e_{it} + u_{it}$ as an AR(1) process: $w_{it} = \rho w_{i,t-1} + \epsilon_{it}$ with $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$, and then considered θ_{it} as

$$\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i. \tag{3.8}$$

Tiller (1989) used a similar approach in the context of labour force estimation from aggregate time series data generated from repeated surveys. The combined model, under the above assumption, may be written as

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + v_i + w_{it}, \\ w_{it} &= \rho w_{i,t-1} + \epsilon_{it}, \ |\rho| < 1, \end{aligned} \tag{3.9}$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$ and $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$.

4. EMPIRICAL BEST LINEAR UNBIASED PREDICTOR

4.1 BLUP

Arranging the data $\{y_{it}\}$ as $\mathbf{y} = (y_{11}, \ldots, y_{1T}; \ldots; y_{I1}, \ldots, y_{IT})' = (\mathbf{y}_1', \ldots, \mathbf{y}_I')'$, the model (3.9) can be expressed as a special case of the general mixed model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{w} \tag{4.1}$$

with

$$\mathbf{X}' = (\mathbf{X}_1', \ldots, \mathbf{X}_I')$$
$$\mathbf{Z} = \mathbf{I} \otimes \mathbf{1}_T, \ \beta = (\beta_0, \beta_1)',$$

where $\mathbf{X}_i$ is a $T \times 2$ matrix with t-th row given by $(1, x_{it})$, $\mathbf{I}$ is the identity matrix of order I and $\mathbf{1}_T$ is the t-vector of 1's. Further,

$$E(\mathbf{v}) = 0, \quad Cov(\mathbf{v}) = \sigma_v^2\mathbf{I}$$
$$E(\mathbf{w}) = 0, \quad Cov(\mathbf{w}) = \sigma^2(\mathbf{I} \otimes \Gamma) = \sigma^2\mathbf{R} \text{ (say)}$$

and Γ is a $T \times T$ matrix with (i,j)-th element $\gamma_{ij} = (1-\rho^2)^{-1}\rho^{|i-j|}$.

Henderson (1975) derived the best linear unbiased predictor (BLUP) of any linear combination of β and the random effects $\mathbf{v}$, say $\tau = \mathbf{k}'\beta + \mathbf{m}'\mathbf{v}$, as

$$\tilde{\tau} = \mathbf{k}'\tilde{\beta} + \mathbf{m}'\mathbf{Z}'\Sigma^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta})(\sigma_v^2/\sigma^2). \tag{4.2}$$

Here, $\Sigma = \mathbf{I} \otimes [(\sigma_v^2/\sigma^2)\mathbf{J} + \Gamma]$ with $\mathbf{J}$ denoting a $T \times T$ matrix of 1's, and $\tilde{\beta} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Sigma^{-1}\mathbf{y})$ is the generalized least squares estimator of β. If $\tau = \theta_{it}$ as given by (3.8), then

$$\mathbf{k}' = (1, x_{it}), \ \mathbf{m}' = (0, \ldots, 0, 1, 0, \ldots, 0) \tag{4.3}$$

with 1 in the ith position, and

$$\mathbf{m}'\mathbf{Z}'\Sigma^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta}) = \mathbf{1}_T'[(\sigma_v^2/\sigma^2)\mathbf{J} + \Gamma]^{-1}(\mathbf{y}_i - \mathbf{X}_i\tilde{\beta}). \tag{4.4}$$

4.2 Estimation of σ^2

The BLUP (4.2) depends on the unknown variance ratio σ_v^2/σ^2 and the unknown autocorrelation ρ. We used the method of Pantula and Pollack (1985) to estimate the parameters σ^2, σ_v^2 and ρ. This method is an extension of the method of fitting constants for the case $\rho = 0$ (Fuller and Battese, 1973), and the estimates of σ^2, σ_v^2 and ρ are obtained as below.

Lt $\{\tilde{e}_{it}\}$ be the ordinary least squares residuals obtained by regressing y_{it} on x_{it}, with the intercept term included. Then ρ is estimated by

$$\hat{\rho} = \left[\sum_{i=1}^{I}\sum_{t=1}^{T-2} \tilde{e}_{it}(\tilde{e}_{i,t+1} - \tilde{e}_{i,t+2})\right]\left[\sum_{i=1}^{I}\sum_{t=1}^{T-2} \tilde{e}_{it}(\tilde{e}_{it} - \tilde{e}_{i,t+1})\right]^{-1}. \tag{4.5}$$

Define

$$z_{it}^{(1)} = z_{it} - z_{it}^{(2)},$$

where

$$z_{it} = y_{it} - \hat{\rho}y_{i,t-1}, \quad t \geq 2$$
$$= f_1 y_{it}, \quad t = 1$$

and

$$z_{it}^{(2)} = c^{-1}d_i f_t$$

with

$$c = (1-\hat{\rho})[T - (T-2)\hat{\rho}],$$
$$f_t = 1 - \hat{\rho}^2, \quad t = 1$$
$$= 1 - \hat{\rho}, \quad t \geq 2$$

and

$$d_i = \sum_{t=1}^{T} f_t z_{it}.$$

Similarly define $(h_{0it},\ h^{(1)}_{0it},\ h^{(2)}_{0it})$ and $(h_{1it},\ h^{(1)}_{1it},\ h^{(2)}_{1it})$ in terms of the elements 1 and x_{it}, i.e., replace y_{it} with 1 and x_{it} respectively in the expressions for $(z_{it},\ z^{(1)}_{it},\ z^{(2)}_{it})$. Let $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ be the residual sum of squares obtained by regressing $z^{(1)}_{it}$ on $h^{(1)}_{0it}$ and $h^{(1)}_{1it}$, without the intercept term. Also, define

$$g_i = \sum_{t=1}^{T} f_t z_{it},$$
$$f_{0i} = \sum_{t=1}^{T} f_t h_{0it},\ f_{1i} = \sum_{t=1}^{T} f_t h_{1it}.$$

Let $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ be thre residual sum of squares obtained by regressing g_i on f_{0i} and f_{1i}, without the intercept term. The estimates of σ^2 and σ_v^2 are now obtained as

$$\hat{\sigma}^2 = [I(t-1)-2]^{-1}\hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{4.6}$$

and

$$\hat{\sigma}_v^2 = c^{-1}(I-2)^{-1}[\hat{\mathbf{u}}'\hat{\mathbf{u}} - \hat{\sigma}^2(I-2)], \tag{4.7}$$

in the case of model (3.9).

If $p-1(\geq 2)$ x-variables are included in the model, then $\{\tilde{e}_{it}\}$ are obtained by regressing y_{it} on $x_{1it},\ldots,x_{p-1,it}$, with the intercept term included. Similarly, $(h_{jit},\ h^{(1)}_{jit},\ h^{(2)}_{jit})$, $j = 0,1,\ldots,p-1$ are defined in terms of the elements 1, $x_{1it},\ldots,x_{p-1,it}$, and $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is obtained by regressing $z^{(1)}_{it}$ on $h^{(1)}_{0it},\ h^{(1)}_{1it},\ldots,h^{(1)}_{p-1,it}$ without the intercept term, and $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ is obtained by defining $f_{ji} = \sum_{t=1}^{T} f_t h_{jit}$, $j = 0,1,\ldots,p-1$ and regressing g_i on $f_{0i}, f_{1i},\ldots,f_{p-1,i}$ without the intercept term. Finally, $\hat{\sigma}^2$ and $\hat{\sigma}_v^2$ are defined by (4.6) and (4.7) respectively, with $I(T-1)-2$ changed to $I(T-1)-p$ and $I-2$ changed to $I-p$. . It is also possible to get maximum likelihood estimates of σ^2, σ_v^2 and ρ, using the EM algorithm (see Chi and Reinsel, 1989).

Substituting the estimates $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ and $\hat{\rho}$ in (4.2), we get the empirical best linear unbiased predictor (EBLUP) of θ_{it}, denoted by $\hat{\theta}_{it}$.

5. MEAN SQUARE ERROR OF EBLUP

Following Henderson (1975), the mean square error (MSE) of BLUP, $\tilde{\tau} = \tilde{\theta}_{it}$, is given by

$$\begin{aligned}\text{MSE}\ (\tilde{\theta}_{it}) = \sigma^2\{&\mathbf{k}'(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{k} + (\sigma_v^2/\sigma^2)\mathbf{m}'\mathbf{m} - (\sigma_v^2/\sigma^2)^2\mathbf{m}'\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{Z}\mathbf{m}\\ &-2(\sigma_v^2/\sigma^2)\mathbf{k}'(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Z}\mathbf{m}\},\end{aligned} \tag{5.1}$$

where $\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}$. The MSE of EBLUP, $\hat{\theta}_{it}$, involves lower order terms that take account of the uncertainty in the estimators $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ and $\hat{\rho}$. We are currently developing an accurate approximation to the MSE of EBLUP, along the lines of Prasad and Rao (1990) for the Fay-Herriot model.

In this paper we ignored the uncertainty in the estimators $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ and $\hat{\rho}$, and used (5.1) with $(\hat{\sigma}^2,\ \hat{\sigma}_v^2,\ \hat{\rho})$ substituted for $(\sigma^2, \sigma_v^2,\ \rho)$ as an estimator of MSE of the EBLUP. This estimator underestimates the true MSE of EBLUP, but the underestimation is not likely to be serious for our empirical study in Section 7.

The MSE of the survey estimator, y_{it}, of θ_{it} under the model (3.9) is given by

$$\text{MSE}(y_{it}) = E(y_{it} - \theta_{it})^2 = V(w_{it}) = \sigma^2/(1-\rho^2). \tag{5.2}$$

An estimator of MSE (y_{it}) is obtained by substituting $(\hat{\sigma}^2,\ \hat{\rho})$ for $(\sigma^2,\ \rho)$ in (5.2).

6. SYNTHETIC ESTIMATORS

If we ignore the random small area effects $\{v_i\}$, and use the model

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + w^*_{it},\\ w^*_{it} &= \rho^* w^*_{i,t-1} + \epsilon^*_{it}, \quad |\rho^*| < 1,\end{aligned} \tag{6.1}$$

where $\epsilon^*_{it} \sim_{\text{ind}} N(0, \sigma^{*2})$, we get a synthetic estimator of $\theta_{it} = \beta_0 + \beta_1 x_{it}$. It is given by

$$\tilde{\theta}_{it}(S) = \tilde{\beta}_0(S) + \tilde{\beta}_1(S)x_{it}, \tag{6.2}$$

where $\tilde{\beta}_0(S)$ and $\tilde{\beta}_1(S)$ are the generalized least squares estimators of β_0 and β_1 under the model (6.1): $\tilde{\beta}(S) = (\mathbf{X}'\mathbf{R}^{*-1}\mathbf{X})^{-1})(\mathbf{X}'\mathbf{R}^{*-1}\mathbf{y})$, where $\mathbf{R}^*$ is given by $\mathbf{R}$ with ρ^* substituted for ρ. The estimator $\tilde{\theta}_{it}(S)$ is unbiased for $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i$ under the model of interest, (3.9).

Writing $\tilde{\theta}_{it}(s)$ as a linear function, $\mathbf{a}'\mathbf{y}$, of the observations $\mathbf{y}$, the MSE of $\tilde{\theta}_{it}(S)$ under the model of interest (3.9) can be obtained. It is given by

$$\begin{aligned} \text{MSE}[\tilde{\theta}_{it}(S)] &= E(\mathbf{a}'\mathbf{y} - \mathbf{k}'\boldsymbol{\beta} - \mathbf{m}'\mathbf{v})^2 \\ &= \sigma^2[(\mathbf{Z}'\mathbf{a} - \mathbf{m})'(\mathbf{Z}'\mathbf{a} - \mathbf{m})(\sigma_v^2/\sigma^2) + \mathbf{a}'\mathbf{R}\mathbf{a}], \end{aligned} \tag{6.3}$$

where $\mathbf{k}$ and $\mathbf{m}$ are given by (4.3).

Since $\tilde{\theta}_{it}(S)$ depends on the unknown autocorrelations ρ^*, we estimate ρ^* from (6.1) using the modified Gauss-Newton method (Hartley, 1961). The MSE of the resulting estimator, $\hat{\theta}_{it}(S)$, is estimated by substituting $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})$ for $(\sigma^2, \sigma_v^2, \rho)$ in (6.3). This estimator will underestimate the true MSE of $\hat{\theta}_{it}(S)$ since the uncertainty in estimating ρ^* is ignored.Nevertheless, the underestimation is not likely to be serious for our empirical study in Section 7.

Another synthetic estimator is obtained by considering the fixed effects model (3.2) on θ_{it} and then writing

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + v_i + \tilde{w}_{it} \\ \tilde{w}_{it} &= \tilde{\rho}\tilde{w}_{i,t-1} + \tilde{\epsilon}_{it} \quad |\tilde{\rho}| < 1, \end{aligned} \tag{6.4}$$

where $\tilde{\epsilon}_{it} \sim_{\text{ind}} N(0, \tilde{\sigma}^2)$ and $\{v_i\}$ are fixed small area effects. The resulting synthetic estimator of θ_{it} is given by (Choudhry and Hunter, 1987):

$$\tilde{\theta}_{it}(S1) = \tilde{\beta}_0(S1) + \tilde{\beta}_1(S1)x_{it} + \tilde{v}_i(S1), \tag{6.5}$$

where $\tilde{\gamma}(S1) = [\tilde{\beta}_0(S1),\ \tilde{\beta}_1(S1),\ \tilde{v}_1(S1), \ldots, \tilde{v}_I(S1)]'$ is the generalized least squares estimator given by $(\mathbf{W}'\tilde{\mathbf{R}}^{-1}\mathbf{W})^-(\mathbf{W}'\tilde{\mathbf{R}}^{-1}\mathbf{y})$. Here the $(i,\ t)$-th row of $\mathbf{W}$ is the $1 \times (I+2)$ vector $(1, x_{it}, 0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the $(i+2)$-th position, $\tilde{\mathbf{R}}$ is given by $\mathbf{R}$ with $\tilde{\rho}$ substituted for ρ, and $(\mathbf{W}'\tilde{\mathbf{R}}^{-1}\mathbf{W})^-$ is a generalized inverse of $\mathbf{W}'\tilde{\mathbf{R}}^{-1}\mathbf{W}$. The estimator $\tilde{\theta}_{it}(S1)$ is unique for any choice of generalized inverse.

Writing $\tilde{\theta}_{it}(S1)$ as $\mathbf{b}'\mathbf{y}$, it is seen that $\tilde{\theta}_{it}(S1)$ is biased for θ_{it} under the model of interest (3.9). Its MSE under the model of interest (3.9) is given by

$$\begin{aligned} \text{MSE}[\tilde{\theta}_{it}(S1)] &= E(\mathbf{b}'\mathbf{y} - \mathbf{k}'\boldsymbol{\beta} - \mathbf{m}'\mathbf{v})^2 \\ &= [(\mathbf{X}'\mathbf{b} - \mathbf{k})'\boldsymbol{\beta}]^2 + \sigma^2[(\mathbf{Z}'\mathbf{b} - \mathbf{m})'(\mathbf{Z}'\mathbf{b} - \mathbf{m})(\sigma_v^2/\sigma^2) + \mathbf{b}'\mathbf{R}\mathbf{b}] \end{aligned} \tag{6.6}$$

Since the estimator $\tilde{\theta}_{it}(S1)$ depends on the unknown autocorrelation $\tilde{\rho}$, we estimate $\tilde{\rho}$ from (6.4) using the modified Gauss-Newton method. The MSE of the resulting estimator, $\hat{\theta}_{it}(S1)$, is estimated by replacing $\boldsymbol{\beta}$ with $\tilde{\boldsymbol{\beta}}$ in (6.6) and then substituting $(\hat{\sigma}^2,\ \hat{\sigma}_v^2,\ \hat{\rho})$ for $(\sigma^2,\ \sigma_v^2,\ \rho)$. This estimator will underestimate the true MSE of $\hat{\theta}_{it}(S1)$ since the uncertainty in estimating $\tilde{\rho}$ is ignored. Nevertheless, the underestimation is not likely to be serious for our empirical study in Section 7.

7. EMPIRICAL STUDY

The efficiencies of the EBLUP, and the two synthetic estimators and the survey estimator y_{it} are now evaluated, using 36 months (January '83 - December '85) of survey estimates of unemployment from the Canadian Labour Force Survey for 21 census divisions (small areas) in the province of British Columbia. The auxiliary variables used in the regression are monthly administrative counts from the UI system and the population in the labour force from the Labour Force Survey. Here, letting $t = 1, \ldots, 36$ and $i = 1, \ldots, 21$, y_{it} = log (survey estimate of proportion of population unemployed), x_{1it} = log (Unemployment Insurance beneficiaries/projected population 15 years and over), x_{2it} = survey estimate of labour force participation rate. The labour force participation rate is defined as the proportion of target proportion which is either employed or unemployed. Although x_{2it} is subjected to sampling errors, its coefficient of variation (cv) is negligible compared to the cv of y_{it} and hence these errors may be ignored without affecting the estimates.

Our model (3.9) with two concomitant variables may be written as

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 c_{2it} + v_i + w_{it}$$
$$w_{it} = \rho w_{i,t-1} + \epsilon_{it}, \quad |\rho| < 1 \tag{7.1}$$

where $v_i \sim_{\text{ind}} N(0, \sigma_v^2)$ and $\epsilon_{it} \sim_{\text{ind}} N(0, \sigma^2)$. The estimated MSE of the EBLUP under (7.1) was computed from (5.1) for each $(i,\ t)$ by substituting the estimates $(\hat{\sigma}^2,\ \hat{\sigma}_v^2,\ \hat{\rho})$ and using $(1,\ x_{1it},\ x_{2it})$ for the t-th row of X_i. These estimates were obtained by using the method of Pantula and Pollock (1985), and are as follows:

$$\hat{\sigma}^2 = 0.0391, \quad \hat{\sigma}_v^2 = 0.0175, \quad \hat{\rho} = 0.362.$$

Turning to the synthetic estimator, the estimated MSE of the synthetic estimator, $\hat{\theta}_{it}(S)$, which ignores the random effects $\{v_i\}$ in (7.1), is obtained for each $(i,\ t)$ from (6.3) by substituting $(\hat{\sigma}^2,\ \hat{\sigma}_v^2,\ \hat{\rho})$. Similarly, the estimated MSE of the synthetic estimate $\ddot{\theta}_{it}(S1)$, which treats $\{v_i\}$ as fixed effects in (7.1), is obtained for each $(i,\ t)$ from (6.6) by replacing β with $\tilde{\beta}$ and then substituting $(\hat{\sigma}^2,\ \hat{\sigma}_v^2,\ \hat{\rho})$ for $(\sigma^2,\ \sigma_v^2,\ \rho)$. Finally, an estimate of MSE of the survey estimate y_{it} is obtained from (5.2) by substituting $(\hat{\sigma}^2,\ \hat{\rho})$ for $(\sigma^2,\ \rho)$.

Denote the estimated efficiency of the EBLUP relative to the synthetic estimator $\hat{\theta}_{it}(S)$ as E_{1it} = est MSE $[\hat{\theta}_{it}(S)]$/ est MSE (EBLUP), the estimated efficiency of the EBLUP relative to the synthetic estimator $\hat{\theta}_{it}(S1)$ as E_{2it} = est MSE $[\hat{\theta}_{it}(S1)]$/est MSE (EBLUP), and the efficiency of EBLUP relative to the survey estimate y_{it} as E_{3it} = est MSE(y_{it})/ est MSE (EBLUP). It should be noted that the MSE's of the synthetic estimates $\hat{\theta}_{it}(S)$ and $\hat{\theta}_{it}(S1)$, and the survey estimate y_{it} are estimated under the model (7.1).

The averages of E_{1it}, E_{2it}, and E_{3it} over thirty-six months are computed as $\overline{E}_{1i} = \Sigma_t E_{1it/36}$, $\overline{E}_{2i} = \Sigma_t E_{2it/36}$, and $\overline{E}_{3i} = \Sigma_t E_{3it/36}$ for each small area i, and these values are reported in Table 1.

It is clear from Table 1 that the EBLUP leads to large gains in average efficiency over the survey estimator, $\overline{E}_{3i}$ ranging from 7.56 to 10.11. The gains in average efficiency of the EBLUP over the synthetic estimator $\hat{\theta}_{it}(S)$ are also substantial, $\overline{E}_{1i}$ ranging from 2.67 to 3.56. The average efficiency of the EBLUP over the synthetic estimator $\hat{\theta}_{it}(S1)$, denoted by $\overline{E}_{2i}$, ranges from 0.82 to 6.45. The over-all average efficiency values are as follows: $\overline{E}_1 = \Sigma\overline{E}_{1i}/21 = 3.02$, $\overline{E}_2 = \Sigma\overline{E}_{2i}/21 = 2.50$ and $\overline{E}_3 = \Sigma\overline{E}_{3i}/21 = 8.61$.

Table 1. Average Monthly Efficiency of the EBLUP Relative to the Synthetic Estimators and the Survey Estimator, Under the Model (7.1)

Small Area	$\overline{E}_{1i}$	$\overline{E}_{2i}$	$\overline{E}_{3i}$
1	3.56	1.14	10.11
2	2.86	1.92	8.10
3	2.89	1.63	8.19
4	2.67	3.59	7.56
5	2.87	3.94	8.13
6	3.01	2.87	8.56
7	3.07	0.82	8.72
8	2.98	0.94	8.52
9	3.44	2.05	9.74
10	3.08	1.64	9.72
11	3.24	1.85	9.18
12	2.98	6.45	8.43
13	2.75	1.88	7.80
14	2.98	2.16	8.50
15	3.14	1.98	8.89
16	2.73	4.91	7.74
17	2.76	2.72	7.83
18	2.81	3.37	8.02
19	2.95	4.11	8.34
20	3.43	1.16	9.78
21	3.14	1.44	8.94
Over-all Average	3.02	2.50	8.61

8. CONCLUDING REMARKS

The EBLUP will be derived under the generalized Fay-Herriot model given by (3.7), by first deriving the estimates of the parameters σ^2, σ_v^2 and ρ along the lines of Pantula and Pollack (1985) and then substituting these estimators in the BLUP to get the EBLUP. The efficiency of the EBLUP will be evaluated along the lines of Section 7 using the Canadian Labour Force Survey data and an estimate of the sampling covariance matrix, Σ. Work is in progress on obtaining an estimate of Σ for the Canadian Labour Force Survey.

Accurate approximations to the mean square error of the EBLUP and their estimators will also be drived, along the lines of Prasad and Rao (1990).

REFERENCES

Battese, G.E., Fuller, W.A. and Harter, R. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Binder, D.A. and Dick, J.P. (1989), "Implications of Survey Designs for Estimating Seasonal ARIMA Models," Paper presented at the American Statistical Association Meetings, Washington, D.C.

Brackstone, G.J. (1986), "Small Area Data: Policy Issues and Technical Challenges," in *Small Area Statistics: An International Symposium*, eds, R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh, New York: John Wiley, pp. 3-20.

Chi, E.M. and Reinsel, G.C. (1989), "Models for Longitudinal Data with Random Effects and AR(1) Errors," *Journal of the American Statistical Association*, 84, 452-459.

Choudhry, H. and Hunter, L. (1987), "Time Series Modelling for Samll Area Estimation," in *Statistical Uses of Administrative Data: Proceedings*, Ottawa: Statistics Canada.

Cronkhite, F.R. (1986), "Use of Regression Techniques for Developing State and Area Employment and Unemployment Estimates," *In Samll Area Statistics: An International Symposium*, eds, R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh, New York: John Wiley, pp. 160-174.

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1982), "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey," *Survey Methodology*, 8, 17-47.

Ericksen, E.P. (1974), "A Regression Method for Estimating Population of Local Areas," *Journal of the American Statistical Association*, 69, 867-875.

Fay, R.E. and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Fuller, W.A. and Battese, G.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structures," *Journal of the American Statistical Association*, 68, 626-632.

Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 33-36.

Hartley, H.O. (1961), "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Function by Least Squares," *Technometrics*, 3, 269-280.

Henderson, C.R. (1975), "Best Linear Unbiased Estimation and Prediction under a Selection Model," *Biometrics*, 31, 423- 447.

Jones, R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," *Journal of the Royal Statistical Society*, Ser. B, 42, 221-226.

Judge, G.G., Griffith, W.E., Hill, R.C., Lütkepohl, H., and Lee, T. (1985), *The Theory and Practice of Econometrics*, 2nd ed., New York: John Wiley.

Pantula, S.G. and Pollock, K.H. (1985), "Nested Analysis of Variance with Autocorrelated Errors," *Biometrics*, 41, 909- 920.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of the Mean Square Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85 (in press).

Purcell, N.J. and Kish, L. (1980), "Post censal Estimates for Local Areas (or Domains)," *International Statistical Review*, 48, 3-18.

Rao, J.N.K. (1986), "Synthetic Estimators, SPREE and Best Model-Based Predictors of Small Area Means," in Technical Report No. 97, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Särndal, C.E. and Hidiroglou, M.A. (1989), "Small Domain Estimation: A Conditional Analysis," *Journal of the American Statistical Association*, 84, 266-275.

Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," *International Statistical Reivew*, 45, 13-28.

Tiller, R. (1989), "A Kalman Filter Approach to Labor Force Estimation Using Survey Data," paper presented to the American Statistical Association Meetings, Washington, D.C.

PART 3

ANALYSIS OF TIME SERIES OF COUNTS

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

MAPPING AGGREGATE BIRTH DATA

David R. Brillinger [1]

ABSTRACT

Births by census division are studied via maps for the province of Saskatchewan for the year 1986. A principal goal of the work is to see how births are related to geography by obtaining contour maps displaying the birth phenomenon in a smooth fashion. A hierarchy of models for count-valued random variates are fit to the data by maximum likelihood. Models include: the Poisson, the Poisson with a weekday effect and the Poisson-lognormal. The last mentioned is motivated by the idea that important covariates are unavailable to include in the analysis.

KEY WORDS: Aggregate data; Contouring; Extra-Poisson variation; Locally-weighted analysis; Maps; Poisson distribution; Poisson-lognormal distribution; Random effects; Spatial data; Unmeasured covariates.

1. INTRODUCTION

The concern of this paper is data that has been aggregated over geographical regions. The analysis of such data should be "easy" because of the graphing possibilities, eg. quantity versus geography in the manner of residual plots so often employed in regression analyis; however in the present case the aggregation leads to important difficulties.

The specific data studied consists of daily births for the calendar year 1986 to women aged 25-29 for each of the 18 census divisions of the province of Saskatchewan. The corresponding population sizes, as determined in the 1986 Census, are also employed in order to compute rates. The reason that Saskatchewan was selected for this pilot study is that it is moderate sized and its boundaries and those of its census divisions are regular. (The latter was important at the early stages of the work because computer based maps were unavailable.) Women ages 25-29 were selected because that was the 5 year age group with most births. These data were provided to the author by Statistics Canada.

The data is characterized by being aggregate, by being nonGaussian and by being nonstationary in space and time.

It is wished to understand the relationship of births to geography, specifically to allow spatial patterns of fertility and possible surprises to show themselves. There are two aspects to the study; a locally-weighted analysis of grouped data is developed and random effects models are set down and fit to handle extra-Poisson variation.

It is to be emphasized that this is a preliminary report on work in progress. For example the fine structure of the data is not taken advantage of and no measures of uncertainty of the various estimates have been provided. The paper focuses principally on annual totals for the 18 census divisions. The related paper Brillinger (1990) considers both temporal and spatial aspects.

Saskatchewan has 18 census divisions. These may be seen in Figure 1. That figure also provides the total numbers of births to women aged 25 to 29 for 1986 and the corresponding female population sizes on Census Day, 3 June. (Actually because of Statistics Canada's confidentiality requirements the final digits have been rounded to the nearer of 2 and 7). The small population in the northern half of the province is evident. Figure 2 gives the annual birth rates plotted by census division. The divisions with the lowest values, .131 and .133 births per year, correspond to the cities of Saskatoon and Regina respectively. Figure 3 is a chloropleth map of the rates with intensity of hatching proportional to birth rate.

2. PATCH OR CHLOROPLETH MAPS

Maps of most quantities of direct interest that assign average values to the wholes of counties thereby lie, lie, lie.

In these graphic words Tukey (1979) deplores the use of maps such as those of Figures 2, 3 that are constant across geographic divisions. Indeed examination of Figure 2, as does common knowledge, suggests that the birth phenomenon quite likely varies smoothly across census division boundaries. One of the concerns of this work is to develop maps with smooth variation. It is hoped that such maps will prove useful in the discovery of general models and will allow insightful exploratory analyses.

A second concern is with the statistical distribution of the counts themselves. A natural special stochastic model to employ is the Poisson. Yet the birth process has been found to relate to many socio-economic quantities, eg. diet,

[1] David R. Brillinger, Statistics Department, University of California, Berkeley, CA, 94720

lifestyle, weather, environment, weekday, holidays, age structure. Further the population of the province has varied around the Census Day values throughout 1986 and lastly the women's ages range between 25 and 29. In summary it seems necessary to employ a more flexible model than the Poisson, a model able to handle omitted covariates. The Poisson-lognormal will be employed in this work. As a sideline due to the presence of the standard deviation parameter in the Poisson-lognormal, there will be a borrowing of strength that takes place in combining the data values.

3. LOCALLY-WEIGHTED ANALYSIS

In the case of nonaggregate data, locally-weighted fitting is a convenient fashion by which to estimate smoothly varying quantities. Suppose one has a variate Y with probability distribution $p(Y \mid \theta)$ depending on the finite dimensional parameter θ. Suppose one wishes an estimate of θ particular to the location with coordinates (x,y). Suppose the datum Y_i is available for location (x_i,y_i). Let $W_i(x,y)$ be a weight dependent on the distance of (x_i,y_i) to (x,y).

Consider estimating θ by maximizing the weighted loglikelihood

$$\sum_i W_i(x,y) \log p(Y_i \mid \theta) \tag{1}$$

or (often equivalently) by solving the system of estimating equations

$$\sum_i W_i(x,y)\, \psi(Y_i \mid \hat{\theta}) = 0 \tag{2}$$

with $\psi(Y \mid \theta) = \partial \log p / \partial\theta$, the score function.

To illustrate the technique consider an elementary case, specifically take Y to be normal with mean μ and variance σ^2. The locally weighted estimate of μ results from minimizing

$$\sum_i W_i(x,y)\, [Y_i - \mu]^2$$

and is given by

$$\hat{\mu}(x,y) = \sum_i W_i(x,y)\, Y_i \,/ \sum_i W_i(x,y)$$

an expression with intuitive appeal. It is to be noted that such formulas are commonly used in computer graphics as interpolation procedures, see for example Franke (1982).

Among references we may mention Gilchrist (1967) concerned with "discounting", Pelto *et al.* (1968), concerned with least squares, Cleveland and Kleiner (1975), who suggested the use of moving midmeans and Stone (1977) focusing on regression. In the discussion of Stone's paper, Brillinger (1977) suggested the form (2) for a general distribution and justified it as a Bayes' rule. Cleveland and Devlin (1988) develop the least squares approach in real detail. Tibshirani and Hastie (1987) develop an equi-weighted local likelihood estimation procedure. Staniswalis (1989) studies and implements the general p case. Advantages of the locally-weighted technique include: no "hidden" model distribution assumption, the possibility of discerning nonaditivity, variants for resistance and influence, simple additivity of the observation component, and no matrix inversion (as, for example, kriging requires).

4. CONSTRUCTION OF THE WEIGHTS

The birth data of concern in this work is aggregate (or grouped) totals over census divisions. The procedure of the preceding section cannot therefore be employed directly. The problem is that of obtaining appropriate weights $w_i(x,y)$ evidencing the effect of the census division i on the location (x,y). Suppose $\mid R_i \mid$ denotes the area of census division i. Then the naive weight function is

$$w_i(x,y) = 1/\mid R_i \mid \qquad \text{for } (x,y) \textit{ in } R_i$$

and equal 0 otherwise. In this work functions of the essential form

$$w_i(x,y) = \frac{1}{\mid R_i \mid} \int_{R_i} W(x-u, y-v)\, du\, dv \tag{3}$$

will be employed where $W(.)$ is a kernel appropriate for the nonaggregate case as studied in Cleveland and Devlin (1988). The formula (3) may be motivated by consideration of the Poisson point process case. Estimates will be determined via the criteria (1) or (2) with W_i replaced by w_i.

The specific weights employed at $\mathbf{r} = (x,y)$ are

$$w_i(\mathbf{r}) = \exp\{-(1-\rho)^2 \mid\mid \mathbf{r} - \mathbf{r}_i \mid\mid^2 / 2\tau^2\} \tag{4}$$

outside the ellipse $(\mathbf{r}_0 - \bar{\mathbf{r}}_i)\mathrm{S}^{-1}_i(\mathbf{r}_0 - \mathbf{r}_i)' = d_0^2 = 5.991$ and equal 1 inside. Here $\mid\mid \mathbf{r} \mid\mid^2 = x^2 + y^2$, $\rho = d_0 / \sqrt{(\mathbf{r} - \bar{\mathbf{r}}_i)\mathrm{S}_i^{-1}(\mathbf{r} - \bar{\mathbf{r}}_i)'}$ and $\tau = .025$, where $\bar{\mathbf{r}}_i = E\ U_i$ and $\mathrm{S}_i = var\ U_i$ with U_i a variate uniformly distributed within R_i. The logic is that the census divisions are approximated by ellipses with the same mean and variance-covariance matrix. (The specific values were chosen after a bit of experimentation, in part to make the area in the initial ellipse about .95 of the division's.)

Figure 4 displays the .50 and .99 contours of the $w_i(x,y)$ plotted for several of the census divisions. The contours are seen to follow the general shapes of the census divisions.

Other weight functions constructed with similar problems in mind may be found in Tobler (1979) and Dyn and Wahba (1982). Advantages of the present approach, as listed for the nonaggregate case above include: the terms are additive and do not interact, no matrix inversion is needed, and resistance to outliers is easily built in.

Cliff and Ord (1975) Section 5.1, discusses measures of the influence of counties on each other. The concern of this present paper is the influence of a "county" on a point location.

5. THE SIMPLE POISSON

Throughout the analysis, the female population aged 25-29 and births to its members will be considered. Let $i = 1,..., 18$ index census division. Let N_i denote the census count of the women in the i-th division. (These are the counts for Census Day, 3 June 1986.) Let B_i denote the total number of births to women aged 25-29 in the year 1986.

Suppose that the probability distribution $p(.)$ of Section 3 is that B_i is Poisson with mean $N_i \mu$. The parameter μ is a birth rate. One logic for the Poisson assumption comes from the idea that birthdays are random, see Brillinger (1986).

With the Poisson assumption, the locally weighted estimate of the birth rate at location (x,y) is

$$\hat{\mu}(x,y) = \sum_i w_i(x,y)\, B_i \;/ \sum_i w_i(x,y)\, N_i \qquad (5)$$

These values are computed for (x,y) on a 40 by 40 grid. The corresponding contour plot is given in Figure 5. The contours are seen to vary smoothly. This (smoothed) rate varies from .14 to .20, with the higher values in the upper half of the province and the lower centred around the most urban part of the province.

6. THE POISSON WITH WEEKDAY EFFECTS

While the focus of this paper is on spatial analysis, it is usefult to briefly take some definite note of the temporal aspects that are present. It is common knowledge that birth rates vary with the day of the week due to medical intervention, see for example Miyaokoa (1989). The total number of births cannot therefore be reasonably expected to be a homogeneous Poisson. The following model seems worth considering. Let j be an indicator variable with $j = 1$ if the measurement is for a weekday and $j = 2$ if the measurement is for a weekend. Let B_{ij} denote the corresponding number of births in census division i. Suppose that B_{ij} is Poisson with mean $N_i \exp\{\alpha+\beta_j\}$. β_j is the weekday effect and it will be assumed that $\beta_1 + \beta_2 = 0$ to make the model identifiable. If there is no weekday effect, then $\beta_1, \beta_2 = 0$. Now, via locally-weighted estimation as described in Sections 3 and 4, one can obtain estimates of α and β as functions of location.

Figure 6 provides the estimate $\exp\{\hat{\alpha}(x,y)\}$ obtained of the annual birth rate. It is interesting to note that, relative to the constant Poisson model, the contours have expanded out from the urban area for the annual rate. Figure 7 provides the estimated weekday effect $\hat{\beta}_1(x,y)$. In its case there is bulge to the east. The order of magnitude of the $\hat{\beta}$'s is .00 to .10 while $\hat{\alpha}$ is order -2.0 to -1.6 .

The just preceding analysis suggests that there are basic variables that can affect birth rates and that modelling and analysis needs to take this circumstance into account.

7. THE POISSON-LOGNORMAL

With a multi-dimensional explanatory variable $\mathbf{x}_i$ in hand, a Poisson model that has B_i of mean $N_i \exp\{\mathbf{x}_i \theta\}$ might do a good job of explaining the data. Examples of explanatory variables include: diet, lifestyle, weather, environment, holidays, population change, age structure, vagaries of boundaries. In the present situation, these variables are not at hand. The omitted variables in the model will be assumed specifically accumulated into an error variable. It will be assumed that, given ε_i, the variate B_i is Poisson with mean $N_i \mu \exp\{\varepsilon_i\}$ and that ε_i is normal with mean 0 and variance σ^2. Here B is said to have a Poisson lognormal distribution. Some information on this distribution may be found in Shaban (1988).

A central difficulty, that arises in working with a Poisson-lognormal model, is that closed expressions do not exist for the probabilty function. Yet it is clearly flexible for introducing effects and handling missing variables. Following the work of Bock and Lieberman (1970) and Pierce and Sands (1975) however, one can proceed via numerical integration. The probability function may be written

$$p(y) = \frac{1}{y!}\int (ve^{\sigma z})^y \exp\{-ve^{\sigma z}\}\phi(z)dz$$

with ϕ the standard normal density, with y corresponding to B and with v corresponding to $N\mu$. The integral is approximated by a finite number of terms involving nodes and weights.

Figures 8 and 9 provides the result of fitting employing 61 nodes. Figure 8 again shows a dip around the urban region as in Figures 5 and 6. The irregularities suggest that perhaps the estimation procedure converged to a local extremum. Figure 9 is not easily described. It suggests that the estimate is fairly variable. The estimate $\hat{\sigma}$ is seen to be of order of magnitude .1 and so comparable to the weekday effect of Section 6.

8. DISCUSSION

Locally-weighted analysis and random effect models appear to provide a flexible means of dealing with a broad class of problems involving geographic data. The random effect terms have two important roles: handling omitted effects and borrowing strength for improved estimates of the principal parameters. For the Poisson alone, naive totals are efficient, yet there exists extra-Poisson variability due to ommited variables in the present case. The approach is computer intensive, because of the numerical integration and the maximum likelihood estimation at many points on a grid, but proved quite mangeable on the Berkeley network of Sun 3/50's.

Much future work remains including: tools for assessing fit, uncertainty computation, weight function choice (including choice of τ in (4)), analyses for other age groups and provinces, and appropriate asymptotics. Some further results are provided in Brillinger (1990).

Other recent papers devoted to the analysis of vital statistics rates are: Clayton and Kaldor (1987), Tsutakawa (1988) and Manton *et al.* (1989). These papers are not directed at the problem of obtaining a smooth surface, which is the concern of this work.

After the analyses were completed it was learned that the birth counts were based on 1981 census divisions, while the population counts were based on 1986. The boundaries have not changed much, but this provides even more reason for wanting a procedure that can handle extra-variation.

ACKNOWLEDGEMENTS

The author would like to thank G. Brackstone, R. Gussela, R. Raby, B. Sander, P. Spector, M. Subhani, R. Villani for assistance in obtaining the data and maps, for help with computational geometry and with parallel computing. The research was supported by National Science Foundation Grant DMS-8900613.

REFERENCES

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika 35, 179-197.

Brillinger, D. R. (1977). Discussion of Stone (1977). Ann. Statist. 5, 622-623.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. Biometrics 42, 693-734.

Brillinger, D. R. (1990). Spatial-temporal modelling of spatially aggregate birth data. Tech. Report, Statistics Dept., University of California, Berkeley.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83, 596-610.

Cleveland, W. S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. Technometrics 17, 447-454.

Cliff, A. D. and Ord, J. K. (1975). Model building and the analysis of spatial pattern in human geography. J. Royal Stat. Soc. 37, 297-348.

Dyn, N. and Wahba, G. (1982). On the estimation of functions of several variables from aggregated data. SIAM J. Math. Anal. 13, 134-152.

Franke, R. (1982). Scattered data interpolation: tests of some methods. Math. Comp. 38, 181-200.

Gilchrist, W. G. (1967). Methods of estimation involving discounting. J. Royal. Stat. Soc. 29, 355-369.

Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pelom, A. C. (1989). Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates. J. Amer. Statist. Assoc. 84, 637-650.

Miyaoka, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. Canadian J. Stat. 17, 123-140.

Pelto, C. R., Elkins, T.A. and Boyd, H.A. (1968). Automatic contouring of irregularly spaced data. Geophysics 33, 424-430.

Pierce, D. A. and Sands, B. R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Dept., Oregon State University.

Preparata, F. P. and Shamos, I. (1985). Computational Geometry. Springer, New York.

Shaban, S. A. (1988). Poisson log-normal distributions. Pp. 195-210 in Lognormal Distributions (eds. E. L. Crow and K. Shimizu). M. Dekker, New York.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. J. Amer. Statist. Assoc. 84, 276-283.

Stone, C. J. (1977). Consistent nonparametric regression. Ann. Statist. 5, 595-620.

Tibshirani, R. and Hastie, T. 91987). Local likelihood estimation. J. Amer. Statist. Assoc. 82, 559-567.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. J. Amer. Statist. Assoc. 74, 519-536.

Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. J. Amer. Statist. Assoc. 83, 37-42.

Tukey, J. W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Included in The Collected Works of J. W. Tukey, Vol. 5 (1988), (Ed. W. S. Cleveland). Wadsworth, Pacific Grove.

APPENDIX

In this Appendix a few computing details are provided. The census divisions and the province boundaries are specified as polygons. To compute the weights $w_i(x,y)$ a routine was required to check whether a given point was inside a given polygon. To compute the mean and variance of a random point inside a given polygon, a procedure breaking the polygon up into triangles was required. Such routines are discussed in Preparata and Shamos (1985). The likelihood was maximized via the Harwell FORTRAN routine va09a. For the parallel computations the 40 by 40 grid was broken up into 20 disjoint segments.

FIGURE LEGENDS

Figure 1. Births for the 18 census divisions of Saskatchewan for the year 1986 to women in the 25-29 age group and corresponding total numbers of women in that age group on June 3 of the year. (As discussed in the text, the final digits of counts have been rounded to the nearer of 2 and 7.)

Figure 2. Annual birth rates for the 18 census divisions for women aged 25 to 29.

Figure 3. The rates of Figure 2 displayed via intensity of hatching.

Figure 4. The weights, $W_i(x,y)$ applied in equations (1) or (2) computed via expression (4) for four of the census divisions. They are not shown for all the divisions in the interests of clarity.

Figure 5. Expression (5) graphed for the weights of (4) with B_i the count of births in census division i and N_i the corresponding population count of women aged 25-29.

Figure 6. The estimated birth rate assuming that the number of births, B, given the population at risk, N, is Poisson with mean $N\ \exp\{\alpha \pm \beta\}$ with the plus sign for weekdays and minus for weekends. Local weighted fitting is carried out to obtain the estimate $\exp\{\hat{\alpha}(x,y)\}$.

Figure 7. Plot of the estimated weekday effect $\hat{\beta}(x,y)$ obtained as per Figure 6.

Figure 8. A plot comparable to Figure 6, except that now a normal error term is added to the linear predictor.

Figure 9. A plot comparable to Figure 7, except now (as in Figure 8) a normal error term has been added to the linear predictor.

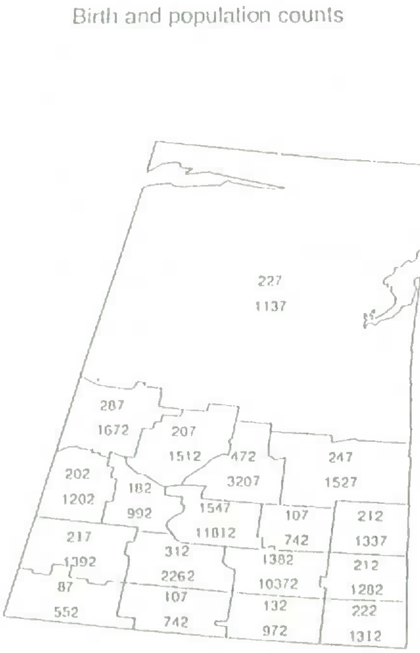


Figure 1

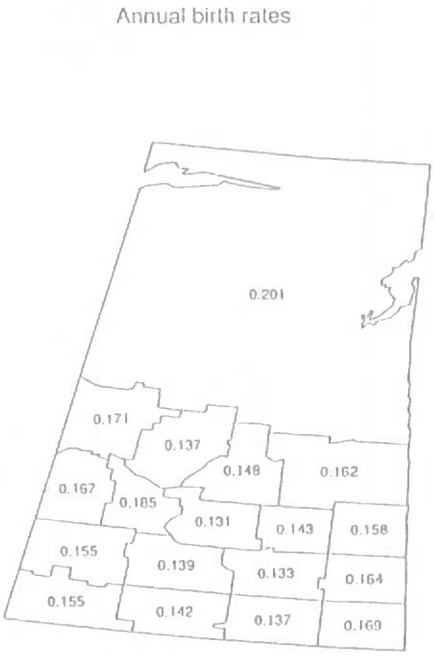


Figure 2

Annual birth rates

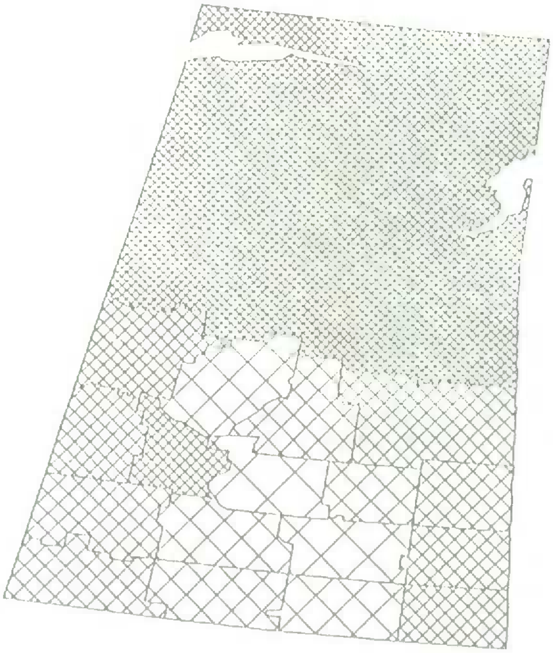


Figure 3

Census division weights

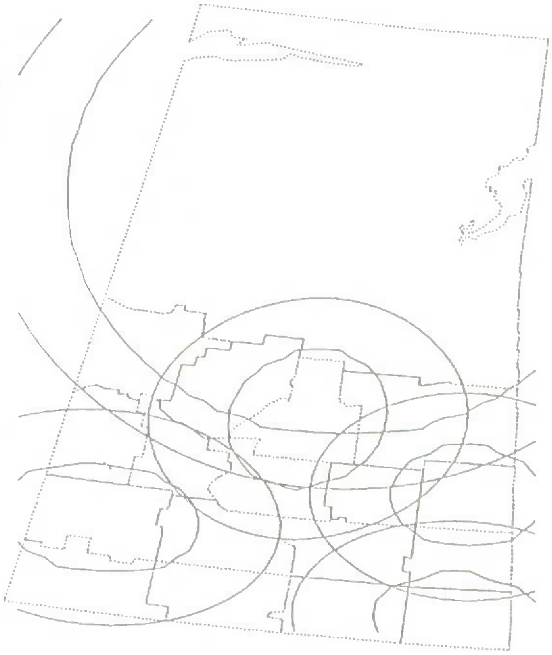
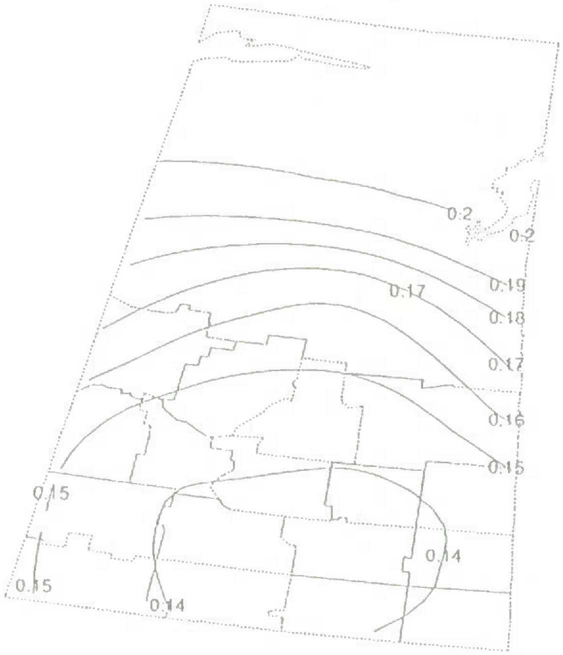


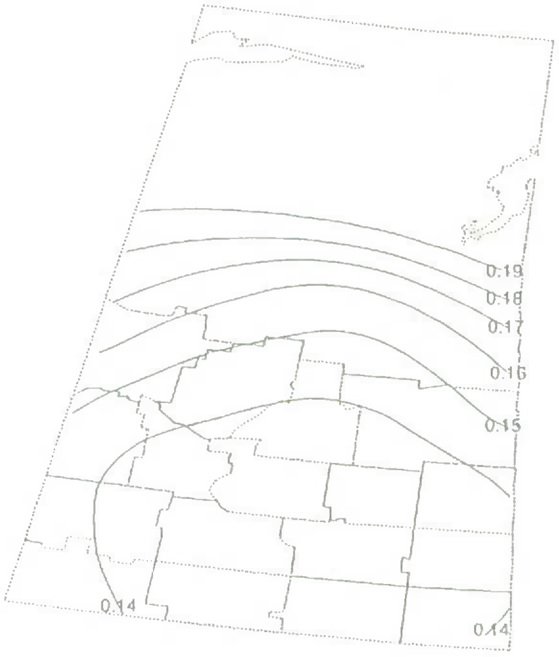
Figure 4

Annual birth rates



Simple Poisson
Figure 5

Annual birth rates

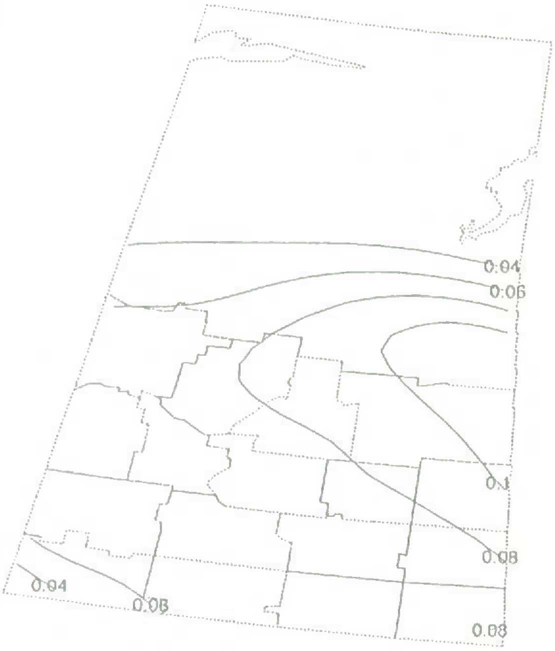


Poisson with weekday effect
Figure 6

Weekday effects

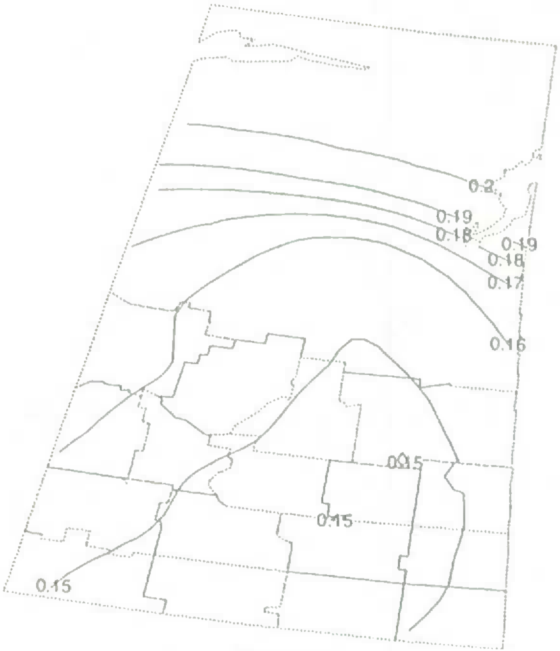
Annual birth rates

Weekday effects



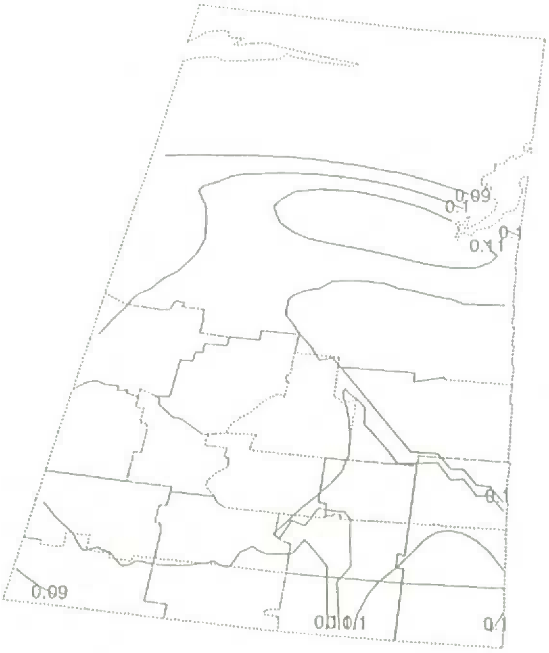
Poisson with weekday effect
Figure 7

Annual birth rates



Poisson-lognormal
Figure 8

Sigma estimates



Poisson-lognormal
Figure 9

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

REGRESSION MODELS FOR PARALLEL TIME SERIES OF COUNTS

Richard Burnett, Daniel Krewski and Jennifer Shedden

ABSTRACT

In this paper, regression models for parallel time series of count data are considered. In particular, we examine the effects of random effects mixing processes to model the variation in response between series, overdispersion within each series, and time dependent correlation. Estimating equations are employed to estimate both the regression and overdispersion parameters.

1. INTRODUCTION

Regression models for count data subject to overdispersion have undergone vigorous development in recent years (McCullagh & Nelder, 1983). Cox (1981) examined models with overdispersion proportional to the variance of the observations, while Breslow (1984), Morton (1987), Lawless (1987) and Dean & Lawless (1989) considered negative binomial type variance structures which arise as a compound Poisson distribution. The compound Poisson-normal and Poisson-Inverse-Gaussian cases have been considered by Hinde (1982) and Dean et al. (1989) respectively, while Brillinger & Preisler (1983) examined arbitrary compound Poisson distributions. Nested random effects models for count data have been studied by Morton (1987) using quasi-likelihood methods, and have been extended to the exponential family by Anderson & Hinde (1988) employing full likelihood methods and the EM algorithm. Zeger et al. (1988) considered a similar problem, but with random effects associated with measured covariates. Autocorelation has been incorporated into models for a single series of count data by Zeger (1988) and Zeger & Qaqish (1988).

In this paper we focus on regression models for parallel time series of count data. Such data arise in the study of the effects of ambient air pollution on daily hospital admission rates for respiratory illnesses (Bates and Sizto, 1987). Since a number of hospitals are usually examined, several time series of the number of daily admissions will be available for analysis. Since hospital records are maintained historically for many years, we consider estimation procedures which are applicable with long time series. A full likelihood approach to parameter estimation usually requires numerical integration or assumptions concerning the magnitude of the overdispersion (Zeger et al., 1988). For long time series, numerical integration can be computationally burdensome. However, Zeger (1988) considered a single time series of counts and employed an estimating equation approach to estimating both regression and overdispersion parameters. No limitations on the degree of overdispersion are needed and the method does not require the use of numerical integration. In this paper, we extend this method to the case of multiple time series.

2. MODEL DEFINITION

Let y_{it} denote the observation on the t^{th} occasion from the i^{th} series $(t = 1, \ldots, n_i;\ i = 1, \ldots, N)$. Although we will assume that the observations are equally spaced in time, missing data can be accommodated in the analysis. Let $\mathbf{x}_{it}$ be a $(p \times 1)$ vector of covariates with an associated vector of regression parameters $\beta = (\beta_1, \ldots, \beta_p)^T$. Further, let ϵ_{it} be a strictly positive random variable with unit expectation, and covariance given by ·

$$\mathrm{Cov}(\epsilon_{it},\ \epsilon_{i,t+\ell}) = \phi\rho_\ell, \tag{2.1}$$

where $\phi > 0$ and $|\rho_\ell| < 1$ represents the lag $\ell = 1, 2, \ldots,\ k \leq \max\{n_i\}$ autocorrelation. Let η_i represent the random effect for the i^{th} series with unit expectation and variance $\tau > 0$, independent of ϵ_{it}. Following Zeger (1988), we assume that the conditional expectation and variance of the observations are defined by

$$E\left(y_{it}|\eta_i,\ \epsilon_{it}\right) = \ \mathrm{Var}\left(y_{it}|\eta_i,\ \epsilon_{it}\right) = \eta_i\ \epsilon_{it}\ \lambda_{it}, \tag{2.2}$$

where $\lambda_{it} = \exp\left(\mathbf{x}_{it}^T\beta\right)$. The conditional covariance between any two observations within the same series is assumed to be zero. The mean, variance and covariance between two observations within the same series, given η_i, are then

$$\begin{aligned} E\left(y_{it}|\eta_i\right) &= \eta_i\lambda_{it} \\ \mathrm{Var}\left(y_{it}|\eta_i\right) &= \eta_i\lambda_{it} + \phi\eta_i^2\lambda_{it}^2, \text{ and} \\ \mathrm{Cov}\left(y_{it},\ y_{i,t+\ell}|\eta_i\right) &= \phi\rho_\ell\eta_i^2\lambda_{it}\lambda_{i,t+\ell}. \end{aligned} \tag{2.3}$$

The unconditional mean, variance and covariance are given by

$$\begin{aligned} E\,(y_{it}) &= \lambda_{it} \\ \mathrm{Var}\,(y_{it}) &= \lambda_{it} + (\tau + \phi[\tau+1])\ \lambda_{it}^2, \text{ and} \\ \mathrm{Cov}\,(y_{it},\ y_{i,t+\ell}) &= (\tau + \phi[\tau+1]\rho_\ell)\,\lambda_{it}\ \lambda_{i,t+\ell}. \end{aligned} \tag{2.4}$$

The variance-covariance matrix for the i^{th} time series is

$$\mathbf{Cov}(\mathbf{Y}_i) = \mathbf{\Lambda}_i + \mathbf{\Lambda}_i\mathbf{R}_i(\alpha)\mathbf{\Lambda}_i \equiv \mathbf{V}_i, \tag{2.5}$$

where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\mathbf{\Lambda}_i =$ *diag* $(\lambda_{i1}, \ldots, \lambda_{in_i})$, $\alpha = (\phi, \tau, \rho_1, \ldots, \rho_k)$ and

$$\mathbf{R}_i(\alpha) = \tau\mathbf{J}_i + \phi(\tau+1)\mathbf{\Omega}_i. \tag{2.6}$$

Here, J_i is a $(n_i \times n_i)$ matrix of ones and Ω_i is $(n_i \times n_i)$ correlation matrix with entries on the ℓ^{th} diagonal given by ρ_ℓ. Our objective is to estimate the regression parameters β and the overdispersion parameters α.

3. PARAMETER ESTIMATION

Since no assumptions have been made concerning the distribution of the conditional observations or the mixing random variables ϵ_{it} and η_i, a likelihood approach to estimation is not possible. However, since the first two moments of the observations have been defined, estimating equations for repeated measures data may be used (Liang & Zeger, 1986).

Given a $N^{1/2}$ consistent estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, the estimate $\hat{\boldsymbol{\beta}}$ of the regression vector $\boldsymbol{\beta}$ satisfies the estimating equation

$$\mathbf{U}(\hat{\beta}|\hat{\alpha}) = \sum_{i=1}^{N} \mathbf{D}_i(\hat{\beta})^T\mathbf{V}_i^{-1}(\hat{\beta},\ \hat{\alpha})\left(\mathbf{Y}_i - \lambda_i(\hat{\beta})\right) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{D}_i(\beta) = \partial\boldsymbol{\lambda}_i/\partial\beta = \mathbf{\Lambda}_i\mathbf{X}_i$, with $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{in_i})^T$ and $\mathbf{X}_i = (x_{i1}, \ldots, x_{in_i})^T$. The estimate $\hat{\beta}$ is determined by an iterative procedure (Liang & Zeger, 1986). Given the current estimate $\hat{\beta}^{(h)}$ of β and $\hat{\alpha}^{(h)}$ of α, the updated estimate $\hat{\beta}^{(h+1)}$ is given by

$$\hat{\beta}^{(h+1)} = \hat{\beta}^{(h)} + \mathbf{H}\left(\hat{\beta}^{(h)}, \hat{\alpha}^{(h)}\right)^{-1}\mathbf{U}\left(\hat{\beta}^{(h)}|\hat{\alpha}^{(h)}\right) \tag{3.2}$$

where

$$\mathbf{H} = -E(\partial\mathbf{U}/\partial\beta) = \sum_{i=1}^{N}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i. \tag{3.3}$$

Moment estimates of ϕ and ρ_ℓ are determined by noting that ϕ represents overdisperion of the observations within a given series and ρ_ℓ represents the lag ℓ autocorrelation. A consistent estimator $(n_i \to \infty)$ of the average conditional variance $\sum_{t=1}^{n_i} \mathrm{Var}(y_{it}|\eta_i)/n_i$ in the i^{th} series is

$$\hat{v} = \sum_{t=1}^{n_i}\left(y_{it} - \hat{\eta}_i\hat{\lambda}_{it}\right)^2/n_i \tag{3.4}$$

where $\hat{\lambda}_{it} = \exp(\mathbf{X}_{it}^T\hat{\beta})$ and

$$\hat{\eta}_i = \left(\sum_{t=1}^{n_i} y_{it}\right)\left(\sum_{t=1}^{n_i}\hat{\lambda}_{it}\right)^{-1} \tag{3.5}$$

is a consistent estimator of η_i. It follows from (2.3) that a consistent estimator $\hat{\phi}$ of ϕ, as $N \to \infty$, is given by

$$\hat{\phi} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}\left[\left(y_{it} - \hat{\eta}_i\hat{\lambda}_{it}\right)^2 - \hat{\eta}_i\hat{\lambda}_{it}\right]}{\sum_{i=1}^{N}\sum_{t=1}^{n_i}\left(\hat{\eta}_i\hat{\lambda}_{it}\right)^2}. \tag{3.6}$$

An estimate for ρ_ℓ is obtained by equating a moment estimate of the lag ℓ autocorrelation to it's expectation, given η_i, which is derived from (2.3), yielding

$$\hat{\rho}_{\ell} = \frac{\sum_{i=1}^{N} \sum_{t=\ell+1}^{n_i} \left(y_{it} - \hat{\eta}_i \hat{\lambda}_{it}\right)\left(y_{i,t-\ell} - \hat{\eta}_i \hat{\lambda}_{i,t-\ell}\right)}{\hat{\phi} \sum_{i=1}^{N} \sum_{t=\ell+1}^{n_i} \hat{\eta}_i^2 \hat{\lambda}_{it} \hat{\lambda}_{i,t-\ell}} \tag{3.7}$$

for $\ell = 1, \ldots, k$. The correlation parameters of an autogressive process are estimated from the $\hat{\rho}_\ell$ by the Yule-Walker equations (Zeger, 1988). For a first order autoregressive process, the correlation parameter is estimated by the lag one autocorrelation $\hat{\rho}_1$.

Finally, a moment estimate of τ is obtained by equating the moment estimate of the unconditional variance to it's expectation, yielding

$$\hat{\tau} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{n_i} \left[\left(y_{it} - \hat{\lambda}_{it}\right)^2 - \hat{\lambda}_{it}\left(1 + \hat{\phi}\hat{\lambda}_{it}\right)\right]}{\left(\hat{\phi} + 1\right) \sum_{i=1}^{N} \sum_{t=1}^{n_i} \hat{\lambda}_{it}^2}. \tag{3.8}$$

The estimation procedure is completed by updating estimates of $\boldsymbol{\beta}$, given by (3.2), and $\boldsymbol{\alpha}$, given by (3.6)-(3.8) until convergence. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\mathbf{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{H}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}\right)^{-1}$. Note that the covariance of $\hat{\boldsymbol{\beta}}$ does not depend on the variance of $\hat{\boldsymbol{\alpha}}$ due to the independence of the unconditional expectation λ_{it} and the overdispersion parameters $\boldsymbol{\alpha}$.

Since only the first two moments of η_i or ϵ_{it} have been specified, estimates of the error in $\hat{\boldsymbol{\alpha}}$ are not, in general, available. However, the focus of our analysis is on the regression parameters $\boldsymbol{\beta}$, with overdispersion treated as a nuisance factor. We suggest a sensitivity analysis on $\mathbf{Cov}\left(\hat{\boldsymbol{\beta}}\right)$ with respect to the form of $\boldsymbol{\alpha}$ as a means of selecting an appropriate overdispersion model. For example, setting $\tau = 0$ indicates that the parallel series may be considered as a single aggregate time series. If $\rho = 0$, then the observations will not display an autoregressive error structure. If $\tau = \rho = 0$ and $\phi > 0$, then the data represent a single series of uncorrelated overdispersed counts. Other approaches to detecting overdispersion in a single series of counts have been discussed by Dean et al. (1989).

4. ESTIMATION FOR LONG TIME SERIES

In many applications, long time series are obtained. Our methods require the repeated inversion of the variance-covariance matrix $\mathbf{V}_i$, which, for protracted series, can be computationally burdensome. We circumvent this problem by employing a working covariance matrix which can be algebraically inverted when estimating $\boldsymbol{\beta}$ by (3.2). The dispersion matrix of the vector of regression parameter estimates is then calculated using the actual covariance. This approach has been employed for a single series of counts by Zeger (1988). Consider first an approximation $\widetilde{\mathbf{V}}_i$ to $\mathbf{V}_i$ which has the form

$$\widetilde{\mathbf{V}}_i = \widetilde{\boldsymbol{\Lambda}}_i \boldsymbol{\Omega}_i \widetilde{\boldsymbol{\Lambda}}_i + \tau \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T, \tag{4.1}$$

where $\widetilde{\Lambda}_i =$ *diag* $\{(\lambda_{i1} + \phi(\tau+1)\lambda_{i1}^2)^{1/2}, \ldots, (\lambda_{in_i} + \phi(\tau+1)\lambda_{in_i}^2)^{1/2}\}$. Note that $\widetilde{\mathbf{V}}_i$ has the same diagonal elements as $\mathbf{V}_i$. Setting $\mathbf{G}_i = \widetilde{\boldsymbol{\Lambda}}_i \Omega_i \widetilde{\boldsymbol{\Lambda}}_i$, it follows from the binomial inverse theorem for matrices (Rao, 1973, p. 33) that

$$\widetilde{\mathbf{V}}_i^{-1} = \mathbf{G}_i^{-1}\left[\mathbf{I} - \tau \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T \mathbf{G}_i^{-1}\left(1 + \tau \boldsymbol{\lambda}_i^T \mathbf{G}_i^{-1} \boldsymbol{\lambda}_i\right)^{-1}\right], \tag{4.2}$$

where $\mathbf{G}_i^{-1} = \widetilde{\boldsymbol{\Lambda}}_i^{-1} \Omega_i^{-1} \widetilde{\boldsymbol{\Lambda}}_i^{-1}$. For an autoregressive autocorrelation structure, $\boldsymbol{\Omega}_i^{-1}$ may be determined explicitly (Zeger 1988). The estimated covariance of the regression parameter estimates employing the working covariance is given by

$$\text{Cov}\ (\tilde{\beta}) = \tilde{\mathbf{H}}^{-1}\left(\sum_{i=1}^{N} \mathbf{D}_i^T \widetilde{\mathbf{V}}_i^{-1} \mathbf{V}_i \widetilde{\mathbf{V}}_i^{-1} \mathbf{D}_i\right) \tilde{\mathbf{H}}^{-1} \tag{4.3}$$

where $\tilde{\mathbf{H}} = \sum_{i=1}^{N} \mathbf{D}_i^T \widetilde{\mathbf{V}}_i^{-1} \mathbf{D}_i$. This form of the covariance does not require $\mathbf{V}_i$ to be inverted.

For very long time series, much over 100 observations, we suggest an even simpler form for the working covariance. Setting $\boldsymbol{\alpha} = 0$ leads to a working covariance given by $\mathbf{A}_i$, which is easily inverted. Further, since the estimate of $\boldsymbol{\beta}$ does not depend on $\boldsymbol{\alpha}$ in this situation, joint iteration between the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is not required.

5. DISCUSSION

Regression models for parallel time series of counts have been described. These models arise in studies of the health effects of ambient air pollution, currently undertaken within the Health Protection Branch. Here, the daily number of respiratory hospital admissions for several hospitals is associated with daily levels of ambient air pollution monitored in the proximity of each hospital. Three sources of overdispersion are considered: between hospital variation in admission rates; within hospital overdispersion; and time dependent correlation.

Overdispersion is modeled by a random effects mixing processes under the assumption that given the random effects, the conditional expectation is equal to the conditional variance. Since only the first two moments of the conditional observations are defined, estimating equation methods are used to estimate both the regression and overdispersion parameters. Consistent estimates of the parameters are obtained and a consistent estimate of the variance of the regression parameters is also derived. However, estimates of the variance of the overdispersion parameters are not available due to the lack of further information on the higher moments. Nonetheless, if the focus of the analysis is on the regression parameters, and overdispersion is considered as a nuisance factor, then this limitation is not critical in practice.

6. REFERENCES

Anderson, D.A. & Hinde, J. (1988). Random effects in generalized linear models and the EM algorithm. *Communication in Statistics* **17**, 3847-3856.

Bates, D.V. & Sizto, R. (1987). Air pollution and hospital admissions in Southern Ontario: the acid summer haze effect. *Environmental Research* **43**, 317-331.

Breslow, N.E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics* **33**, 38-44.

Brillinger, D.R. & Preisler, H.K. (1983). Maximum likelihood estimation in a latent variable problem. In *Studies in Econometrics, Time Series and Multivariate Statistics* (S. Karlin, T. Amemiya, L.A. Goodman eds.). Academic Press, New York, pp. 31-65.

Cox, D.R. (1981). Statistical analysis of time series, some recent developments. *Scandanivan Journal of Statistics* **8**, 93-115.

Dean, C. & Lawless, J.R. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84**, 467-471.

Dean, C., Lawless, J.R. & Willmot, G.E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics* **17**, 171-181.

Hinde J. (1982). Compounded Poisson regression models. In: GLIM 82 (R. Gilchrist, ed.). Springer, New York, pp. 109-121.

Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**, 209-225.

Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

McCullagh, P. & Nelder, J.P. (1983). *Generalized Linear Models.* Chapman and Hall, New York.

Morton, R. (1987). A generalized linear model with strata of variation. *Biometrika* **74**, 247-257.

Rao, C.R. (1973). *Linear Statistical Inference and its Application.* Wiley, New York.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

ANALYSIS OF CROSS-CLASSIFIED CATEGORICAL TIME SERIES

A.C. Singh and G.R. Roberts[1]

ABSTRACT

A parameter-driven framework for defining generalized linear models for time series data is proposed. The time dependent structure of the cross-sectional parameters is specified through state space models. For this purpose, cross-sectionally consistent estimates of model parameters are utilized. A modification of the Kalman filter, in which the observation vector is suitably transformed, is used in defining the recursive equations for prediction and updating. Application of the proposed method to cross-classified categorical time series of counts is illustrated for the problem of predicting cancer mortality.

KEY WORDS: State space models; Generalized linear models; Kalman filter.

1. INTRODUCTION

The problem of modelling and projecting cross-classified categorical time series is quite common for purposes of planning and policy decisions. The data are generally in the form of a fairly long series of multi-way tables of counts based on a large number of observations collected at regular time intervals. For instance, the Canadian cancer mortality data series represent annual counts for each province cross-classified by cancer site, age and sex; see section 5 for an example. The mortality series are derived from administrative sources with a lag of approximately two years before the data are published. The problem of timeliness has been of major concern among users and researchers and clearly, it would be very useful to project such data series at least up to the current year before their publication. For this purpose, the underlying nature of the data could be considered stochastic (Brillinger, 1986) in spite of their origin from administrative sources. It is then reasonable to assume that there is serial dependence in the series due to certain (known and unknown) common factors. If the data were normal then various familiar time series methods could be employed, see for example the well-known texts by Box and Jenkins (1970), Fuller (1976) and Harvey (1981). However, for non-normal such as Poisson data arising from cancer mortality counts, alternative time series methods should be considered.

There exists considerable research work in analysing non-normal data collected over time. In particular, for repeated categorical outcomes, Koch, Landis, Freeman, Freeman, and Lehnen (1977) use generalized least squares to fit non-linear models in which time is deemed as another classifying factor. The work due to Stiratelli, Laird, and Ware (1984) describes a family of mixed models appropriate for repeated dichotomons responses in which certain assumptions are made about covariance structures. Zeger, Liang, and Self (1985), on the other hand, consider logistic regression models for repeated binary observations under a simple first order auto-regressive time dependence. Methods for modelling ordered categorical outcomes over time are considered by Stram, Wei, and Ware (1988) in which model parameters are assumed to be specific to each occasion or time point and are estimated by maximizing the occasion-specific likelihoods. The joint asymptotic normality of these occasion-specific estimates is used to characterize dependence among repeated observations. The work of Morton (1987) and Preisler (1989) deal with fitting generalized linear models with random effects nested within random day/time effects. The above papers, however, are not concerned with the problem of projection considered in this article.

The time series approaches to non-normal data, can be classified into two types following Cox (1981), namely, observation driven and parameter driven models. Some methods belonging to the former type are due to Kalbfleisch and Lawless (1984, 1985) and Kaufmann (1987) in which Markov models for regression (or transition probabilities) with categorical outcomes are considered; see also Zeger and Qaqish (1988) for a quasi-likelihood approach to Markov regression models for general time series. Another method was recently proposed by Smith and Brunsdon (1989) in which approximate normality is assumed after the multivariate additive-logistic transformation for multinomial data is effected and then ARMA models are employed. Some methods belonging to the latter type, i.e. parameter-driven models, are due to West, Harrison, and Migon (1985) with a Bayesian set up for dynamic extension of generalized linear models, Kitagawa (1987) for a non-normal state space approach in which non-normal densities at each step of the Kalman filter are numerically evaluated, Zeger (1988) with an estimating equation approach where auto-correlation is introduced via a random mixing process, and recently Harvey and Fernandes (1989) with a non-Bayesian state space modelling although conjugate priors are used to specify transition equations.

The above time series methods for non-normal data were developed for univariate data or unidimensional data in the categorical case. While it may be possible to extend these methods to multivariate (or multidimensional) data, the resulting computational requirements seem quite complex. In this article, we propose a simpler alternative when both the number (n_t) of observations at each time point and the total number (T) of time points are reasonably large. Even if T is not large, the estimates of model parameters remain consistent (for n_t large) under fairly mild conditions. The proposed model is termed a state space generalized linear model

[1] A.C. Singh and G.R. Roberts, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A OT6

(SSGLM) in which the technique of Kalman filtering is modified to suit non-normal and non-linear modelling. The modified Kalman filter used in SSGLM is related to the generalized Kalman filter of Zehnwirth (1988) when the link function of SSGLM is identity. The condition n_t large allows for "linearization" of the problem in order to employ the familiar state space linear model methods. This aspect is similar to the transformation idea of Smith and Brunsdon (1989). Also, having n_t large provides consistent cross-sectional parameter estimates which can be conveniently used to specify serial dependence among observations through the transition equation in the state space modelling. This aspect is somewhat related to the approach used in Stram, Wei, and Ware (1988).

Some preliminaries including notation and motivation are first given in section 2. It is seen that the SSGLM formulation arises almost naturally for our problem. In section 3, the proposed method SSGLM is defined within a general set of assumptions similar to those in GLM. Some theoretical results are given in section 4 followed by an illustrative numerical example on projecting cancer mortality data series in section 5. Finally, section 6 contains discussion and suggested directions for future work.

2. PRELIMINARIES

Let $\underset{\sim}{y}_t$ denote the n_t-vector of observations at time t, $t = 1,2 \ldots T$. If the n_t observations are grouped or cross-classified according to some covariates into m domains or groups of interest, then $\underset{\sim}{y}_t$ will also be used to denote the m-vector of estimates, e.g. counts, proportions, or means. It will be assumed that the elements of the vector $\underset{\sim}{y}_t$ for the ungrouped case are independent. However, for the grouped case, they could be dependent. In the following, both n_t and T will be assumed to be large. Symbols "$\sim$" and "$\overset{\cdot}{\sim}$" will be used to denote terms "distributed as" and "asymptotically distributed as" respectively. In the interest of a general framework, we shall work only with the second moment assumptions, i.e. distributions will be specified in the wide sense (WS) only. Suppose

$$\underset{\sim}{y}_t \sim WS\ (\underset{\sim}{\mu}_t,\ \Sigma_t) \tag{2.1a}$$

and

$$\underset{\sim}{y} = (y_1', \ldots, y_T') \sim WS\ (\underset{\sim}{\mu},\ \Omega), \tag{2.1b}$$

where Σ_t is assumed to be nonsingular and may vary with $\underset{\sim}{\mu}_t$. Also, Ω will not, in general, be block diagonal due to serial dependence in the time series of vector observations $\underset{\sim}{y}_t$, $t=1, \ldots T$. Note that if the n_t observations are not grouped, then Σ_t would be a diagonal matrix due to the assumed independence of observations. The problem of interest is to predict $\underset{\sim}{y}_t$ for $t>T$. For this purpose, a suitable model for $\underset{\sim}{\mu}$ as a function of a parsimonious set of parameters $\underset{\sim}{\theta}$ is required such that $\underset{\sim}{\mu}$'s are as close as possible to $\underset{\sim}{y}$'s.

First we define certain notation and terms from Linear Models (LM), Generalized Linear Models (GLM), State Space Linear Models (SSLM), and Random Coefficient Regression (RCR) models. These will be useful in motivating the proposed method described in the next section.

2.1 WLS (Weighted Least Squares) Method from LM Theory

For the cross-section at time t, consider the linear predictor or the model H_{1t}: $\underset{\sim}{\mu}_t = F_t\ \underset{\sim}{\theta}_t$, where $\underset{\sim}{\theta}_t$ is a r-vector of fixed effects ($r \leq m$ for the grouped data case), and F_t is a known covariate (or incidence) matrix. Further assume that Σ_t is constant i.e. does not vary with $\underset{\sim}{\mu}_t$ and is approximately known for large n_t. The optimal estimate $\underset{\sim}{\tilde{\theta}}_t$ of $\underset{\sim}{\theta}_t$ in the Gauss-Markov sense given by the WLS method is obtained as a solution of

$$F_t'\ \Sigma_t^{-1}\ (\underset{\sim}{y}_t - \underset{\sim}{\mu}_t) = 0, \tag{2.2a}$$

which implies that

$$\underset{\sim}{\tilde{\theta}}_t = (F_t'\ \Sigma_t^{-1}\ F_t)^{-1}\ F_t'\ \Sigma_t^{-1}\ \underset{\sim}{\mu}_t. \tag{2.2b}$$

The asymptotic distribution of $\underset{\sim}{\tilde{\theta}}_t$ up to terms of order n_t^{-1} as $n_t \to \infty$ is obtained under a suitable CLT as

$$\underset{\sim}{\tilde{\theta}}_t \overset{\cdot}{\sim} N_r\ (\underset{\sim}{\theta}_t,\ (F_t'\ \Sigma_t^{-1}\ F_t)^{-1}). \tag{2.3}$$

If $Cov(\underset{\sim}{y}_t)$ is known only up to a constant multiple σ^2 of Σ_t, the optimal estimator (2.2b) does not change, but the covariance (2.3) is multiplied by the factor σ^2, the overdispersion parameter.

2.2 IWLS (Iterative Weighted Least Squares) Method from GLM Theory

We next consider the generalization of LM theory to GLM in which Σ_t is allowed to vary with $\underset{\sim}{\mu}_t$ in a known manner and $\underset{\sim}{\mu}_t$ can be a non-linear monotone differentiable function of $\underset{\sim}{\theta}_t$, termed the inverse-link function. The form of the variance-mean relation is motivated from an exponential family distribution which is analogous to the assumption of constant variance motivated from normal distributions. Estimation of $\underset{\sim}{\theta}_t$ for GLM can be carried out by the following method.

Here as before consider t fixed. The linear predictor after transformation through the link function is specified by the model H_{2t}: $g(\underset{\sim}{\mu}_t) = F_t \, \underset{\sim}{\theta}_t$, where $\underset{\sim}{\theta}_t$ is again a r-vector of fixed effects, F_t is a known matrix of covariates and g is the link function. Furthermore, the covariance matrix $\Sigma_t \, (\underset{\sim}{\mu}_t)$ of $\underset{\sim}{y}_t$ is assumed to be a known function of $\underset{\sim}{\mu}_t$. An asymptotically optimal (in the extended Gauss-Markov sense, McCullagh, 1983) estimate $\tilde{\underset{\sim}{\theta}}{}_t^C$ of $\underset{\sim}{\theta}_t$ is given by the solution of the following quasi-likelihood score equation (McCullagh and Nelder, 1989, Ch. 9):

$$D_t' \, \Sigma_t^{-1} \, (\underset{\sim}{Y}_t - \underset{\sim}{\mu}_t) = 0, \tag{2.4}$$

where Σ_t being a function of $\underset{\sim}{\mu}_t$ depends on $\underset{\sim}{\theta}_t$, and D_t is the $n_t$xr matrix $(d\underset{\sim}{\mu}_t/d\underset{\sim}{\theta}_t')$. For the grouped case, D_t would be a m×r matrix. The equation (2.4) can be solved by IWLS based on the Newton-Raphson procedure. For this purpose, first an adjusted dependent variable $\underset{\sim}{z}_t^{(i)}$ is defined for each iteration i, i=1, 2 ... as follows.

$$\underset{\sim}{z}_t^{(i)} = \underset{\sim}{\eta}_t^{(i-1)} + (d\underset{\sim}{\eta}_t/d\underset{\sim}{\mu}_t') \, (\underset{\sim}{y}_t - \underset{\sim}{\mu}_t) \mid_{\underset{\sim}{\theta}_t \, = \, \underset{\sim}{\theta}_t^{(i-1)}}, \tag{2.5}$$

where $\underset{\sim}{\eta}_t = g(\underset{\sim}{\mu}_t)$, and $\underset{\sim}{\mu}_t^{(0)}$ is set equal to $\underset{\sim}{y}_t$. Some ad hoc modification to $\underset{\sim}{y}_t$ may be required if $\underset{\sim}{\eta}_t^{(0)}$ is not well defined. Now for each iteration i, a WLS estimate $\underset{\sim}{\theta}_t^{(i)}$ is obtained using the working model $E \, (\underset{\sim}{z}_t^{(i)}) = \underset{\sim}{\eta}_t^{(i)} = F_t \underset{\sim}{\theta}_t^{(i)}$ along with the working covariance of $\underset{\sim}{z}_t^{(i)}$ given by

$$\Gamma_t^{(i)} = (d\underset{\sim}{\eta}_t/d\underset{\sim}{\mu}_t') \; \Sigma_t (d\underset{\sim}{\eta}_t/d\underset{\sim}{\mu}_t')' \mid_{\underset{\sim}{\theta}_t \, = \, \underset{\sim}{\theta}_t^{(i-1)}} . \tag{2.6}$$

The above process is repeated until convergence. Denoting by $\tilde{\underset{\sim}{\theta}}{}_t^C$ the converged solution, $\underset{\sim}{z}_t$ the corresponding variable from (2.5), and Γ_t the corresponding matrix from (2.6), we have as $n_t \to \infty$,

$$\tilde{\underset{\sim}{\theta}}{}_t^C \; : N_r \, (\underset{\sim}{\theta}_t \, , \, (D_t' \, \Sigma_t^{-1} \, D_t)^{-1}), \tag{2.7}$$

and

$$\underset{\sim}{z}_t \; : WS \, (\underset{\sim}{\eta}_t \, , \, \Gamma_t), \tag{2.8}$$

where (2.7) is valid under a suitable CLT. Note that the length of $\underset{\sim}{z}_t$ increases with n_t in the ungrouped case in which the asymptotic distribution in (2.8) should be interpreted in terms of all finite dimensional marginals of $\underset{\sim}{z}_t$. The above equations (2.7) and (2.8) are GLM analogues of (2.3) and (2.1) respectively in the sense that

$$(D_t' \, \Sigma_t^{-1} \, D_t)^{-1} = (F_t' \, \Gamma_t^{-1} \, F_t)^{-1}, \tag{2.9}$$

because

$$D_t = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\theta}_t') = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}_t')(d\underset{\sim}{\eta}_t'/d\underset{\sim}{\theta}_t') = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}_t') \; F_t, \tag{2.10a}$$

and

$$\Gamma_t^{-1} = (d\underset{\sim}{\mu}_t \; / \; d\underset{\sim}{\eta}_t')' \; \Sigma_t^{-1} \; (d\underset{\sim}{\mu}_t \; / \; d\underset{\sim}{\eta}_t'). \tag{2.10b}$$

Moreover, the estimate $\tilde{\underset{\sim}{\theta}}{}_t^C$ does not change in the presence of overdispersion parameter σ^2 i.e. when $Cov(\underset{\sim}{y}_t)$ is $\sigma^2 \; \Sigma_t \; (\underset{\sim}{\mu}_t)$. The asymptotic variance of $\tilde{\underset{\sim}{\theta}}{}_t^C$ in (2.7), however, changes by a multiplicative factor of σ^2.

Whenever the observation $\underset{\sim}{y}_t$ provides a consistent estimator of $\underset{\sim}{\mu}_t$ (this, for example, would be the case if the n_t observations were grouped into m cells), then an alternative one-step estimator $\underset{\sim}{\theta}_t^{\star}$ can be used following the GSK methodology of Grizzle, Starmer, and Koch (1969). In other words, iteration is stopped after only one cycle and not continued until convergence. The estimator $\underset{\sim}{\theta}_t^{\star}$ can be shown to be asymptotically equivalent to $\tilde{\underset{\sim}{\theta}}{}_t^C$. However, the estimator $\tilde{\underset{\sim}{\theta}}{}_t^C$ would be preferable from finite sample considerations since $\Sigma_t \; (\underset{\sim}{\mu}_t^{(0)})$ may be unstable due to presence of cells with possibly small number of observations. It may be of interest to note that when H_{2t} is a saturated model for grouped data case i.e. when $r = m$, then the two estimators $\tilde{\underset{\sim}{\theta}}{}_t^C$ and $\underset{\sim}{\theta}_t^{\star}$ coincide with each other and are equal to $\underset{\sim}{\theta}_t^{(0)}$ or $F_t^{-1} g(\underset{\sim}{y}_t)$.

2.3 FWLS (Filtered Weighted Least Squares) method from SSLM theory

We now consider the generalization of LM theory to SSLM (state space linear models) in order to allow for serial dependence. In SSLM (see e.g. Harvey, 1981, chapter 4, and Harvey, 1984), the serial dependence is introduced via randomly varying parameters $\underset{\sim}{\theta}_t$, $t=1, \ldots, T$, which are connected by state space models. For the problem considered in this article in which both n_t and T are assumed to be large, it seems natural as well as convenient to attempt modelling $\underset{\sim}{y}_t$ contemporaneously for cross-sectional behaviour and then model the underlying parameters $\underset{\sim}{\theta}_t$ temporally for longitudinal behaviour as in state space modelling, i.e. the model is specified by two equations. In this subsection, estimation methods for SSLM are briefly summarized. Unlike GLM, variance is not allowed to vary with mean, and only the identity link function is used. However, the method proposed in the next section generalizes SSLM in the same way as GLM extends LM in order to provide a suitable method for the problem described earlier in the introduction.

Unlike the previous two subsections, we consider both cross-sectional and longitudinal data together, i.e., the time series of vector observations $\underset{\sim}{y}_t$, $k=1, \ldots T$. The vector $\underset{\sim}{y}_t$, as mentioned earlier, is either a n_t-vector for ungrouped data or a m-vector for the grouped case. Two equations are used for modelling in SSLM, see e.g. Zehnwirth (1988). First, for cross-sectional behaviour, the <u>measurement equation</u> is defined as

$$\underset{\sim}{y}_t = F_t \underset{\sim}{\theta}_t + \underset{\sim}{\varepsilon}_t, \tag{2.11}$$

where F_t is a known matrix of covariates, $\underset{\sim}{\theta}_t$ is a r-vector of random parameters termed the state vector, and the distribution of random errors $\underset{\sim}{\varepsilon}_t$ up to second moments is

$$\underset{\sim}{\varepsilon}_t | \underset{\sim}{\theta}_t \sim WS(0,V_t), \; Cov(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_s | \underset{\sim}{\theta}_t \, \underset{\sim}{\theta}_s) = 0 \text{ for } t \neq s. \tag{2.12}$$

The covariance matrix V_t does not depend on $\underset{\sim}{\theta}_t$ and is assumed to be known for every t. Next, for longitudinal behaviour, the <u>transition equation</u> is defined as

$$\underset{\sim}{\theta}_t = G_t \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\xi}_t, \tag{2.13a}$$

where G_t is a known $r \times r$ transition matrix, and the errors $\underset{\sim}{\xi}_t$ are specified by

$$\underset{\sim}{\xi}_t \sim WS(0,W_t), \; Cov(\underset{\sim}{\xi}_t, \underset{\sim}{\xi}_s) = 0, \; s \neq t, \text{ and}$$

$$Cov(\underset{\sim}{\xi}_t, \underset{\sim}{\varepsilon}_s | \underset{\sim}{\theta}_s) = 0 \text{ for all } s,t; \; Cov(\underset{\sim}{\xi}_t, \underset{\sim}{\theta}_s) = 0 \text{ for } t > s. \tag{2.13b}$$

The covariance matrix W_t is also assumed to be known. It may be noted that the Markov-type assumption in the transition equation (2.13) is made for the purpose of recursive estimation and is not required for optimality considerations.

The model defined by (2.11) and (2.13) is completely specified except for the distribution of the initial state vector $\underset{\sim}{\theta}_0$. Here we shall not consider the usual initialization methods as described in Harvey (1981, Ch. 4) and Harvey and Peters (1984), which is then followed by the optimal estimation of parameters $\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2, \ldots, \underset{\sim}{\theta}_T$ successively by the Kalman filter (KF). Instead, we shall first consider a reduced form of (2.11) and (2.13) into a single equation containing only one parameter vector $\underset{\sim}{\theta}_T$ and then a suitable method of estimating $\underset{\sim}{\theta}_T$ which will be needed for predicting $\underset{\sim}{y}_t$ for $t>T$. This approach will be useful in relating SSLM to LM and GLM described earlier.

Conditional on $\underset{\sim}{\theta}_T$, the model (2.11) and (2.13) can be written as a LM for $\underset{\sim}{y} = (\underset{\sim}{y}_1', \ldots, \underset{\sim}{y}_T')'$ as in Harvey and Peters (1984). Writing $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_{T-1}$ in terms of $\underset{\sim}{\theta}_T$ and $\underset{\sim}{\xi}_t$'s, we get

$$\underset{\sim}{y} = F_T^{\star} \underset{\sim}{\theta}_T + \underset{\sim}{\xi}_T^{\star}, \tag{2.14}$$

where $F_T^{\star}$ is a known $T^{\star} \times r$ matrix of fixed values (the order $T^{\star}$ will be mT for the grouped data case and $n_1 + \ldots + n_T$ for ungrouped data), and $\underset{\sim}{\varepsilon}_T^{\star}$ is a new $T^{\star} \times 1$ error vector with mean zero and covariance matrix Ω_T. The matrix Ω_T can be completely specified in terms of the known matrices V_t, W_t, F_t and G_t. Note that the model (2.14) could have been written conditional on $\underset{\sim}{\theta}_\tau$ at any particular point in time $t=\tau$. Now $\underset{\sim}{\theta}_T$ can be estimated optimally using WLS as in LM by the expression

$$\tilde{\underset{\sim}{\theta}}_T^L = (F_T^{\star\prime} \Omega_T^{-1} F_T^{\star})^{-1} F_T^{\star\prime} \Omega_T^{-1} \underset{\sim}{y}, \tag{2.15}$$

where L stands for the longitudinal data used in estimation.

The above expression involves inversion of Ω_T which would generally have a large dimension; therefore, computational difficulty could arise. One can, however, easily evaluate the above WLS estimate using a KF with a flat prior for the initial state vector because the recursive estimates so obtained are BLUPs (best linear unbiased predictors) or MMSLUEs (minimum mean square linear unbiased estimates); see Harvey (1981, p. 105) and Zehnwirth (1988). A suitable modification to the distribution of the initial state vector will be required if some of the elements in the initial state are stationary, see Harvey and Peters (1984). The recursive algorithm for KF which provides BLUP $\tilde{\underset{\sim}{\theta}}^P_{t|t-1}$ of $\underset{\sim}{\theta}_t$ given $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_{t-1}$ and the updated predictor $\tilde{\underset{\sim}{\theta}}^L_t$ given $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_t$ for every $t \geq 2$, is given by

$$\tilde{\underset{\sim}{\theta}}^P_{t|t-1} = G_t \, \tilde{\underset{\sim}{\theta}}^L_{t-1}, \tag{2.16a}$$

$$\tilde{\underset{\sim}{\theta}}^L_t = \tilde{\underset{\sim}{\theta}}^P_{t|t-1} + K_t(\underset{\sim}{y}_t - F_t \, \tilde{\underset{\sim}{\theta}}^P_{t|t-1}), \tag{2.16b}$$

$$K_t = A_{t|t-1} \, F_t' \, (F_t \, A_{t|t-1} F_t' + V_t)^{-1}, \tag{2.16c}$$

$$A_{t|t-1} = G_t A_{t-1} \, G_t' + W_t, \tag{2.16d}$$

$$A_t = (I - K_t F_t) \, A_{t|t-1}, \tag{2.16e}$$

where $A_{t|t-1}$ is the unconditional error covariance matrix of $\tilde{\underset{\sim}{\theta}}^P_{t|t-1}$ i.e. its MSE, and $A_t = A_{t|t}$ i.e. the MSE of $\tilde{\underset{\sim}{\theta}}^L_t$. It can be seen that the values of $\tilde{\underset{\sim}{\theta}}_1$ and A_1 for starting off the KF (2.16) are $\tilde{\underset{\sim}{\theta}}^C_1$ and $(F_1' \, V_1^{-1} \, F_1^{-1})$ respectively where $\tilde{\underset{\sim}{\theta}}^C_1$ is the cross-sectional WLS estimate (2.2) for fixed $\underset{\sim}{\theta}_1$ with Σ_1 replaced by V_1. The matrix K_t is the Kalman gain at time t. The above algorithm also gives recursively the wide sense distributions of $(\tilde{\underset{\sim}{\theta}}^L_t - \underset{\sim}{\theta}_t)$, $t=1, \ldots T$ in the process of computing $\tilde{\underset{\sim}{\theta}}^L_T$. That is for $t=1, \ldots T$.

$$\tilde{\underset{\sim}{\theta}}^L_t - \underset{\sim}{\theta}_t \sim WS(0, A_t). \tag{2.17}$$

In analogy with the IWLS method used for calculating $\tilde{\underset{\sim}{\theta}}^C_t$ for GLM, the above method of computing $\tilde{\underset{\sim}{\theta}}^L_t$ for SSLM by WLS via Kalman filtering will be referred to in this article as the FWLS method, in order to highlight its relationship with the usual WLS method for LM.

The Kalman filter, in addition to providing various BLUPs, also gives a simple method of calculating error sum of squares for the model (2.14) or (2.11) and (2.13) by means of the one-step ahead prediction residuals $\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}^P_{t|t-1}$ and their MSEs. It follows from the equivalence result (B.2) proved in Harvey and Peters (1984) that for any given $t=\tau$,

$$(\underset{\sim}{y} - F^{*\prime}_{\tau} \, \tilde{\underset{\sim}{\theta}}^L_{\tau})' \, \Omega^{-1}_{\tau} \, (\underset{\sim}{y} - F^{*\prime}_{\tau} \, \tilde{\underset{\sim}{\theta}}^L_{\tau}) = \textstyle\sum_{t=1}^{T} \, (\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}^P_{t|t-1})' B^{-1}_{t|t-1} \, (\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}^P_{t|t-1}), \tag{2.18}$$

where for $t \geq 2$,

$$\tilde{\underset{\sim}{y}}^P_{t|t-1} = F_t \, \tilde{\underset{\sim}{\theta}}^P_{t|t-1}, \; B_{t|t-1} = F_t \, A_{t|t-1} \, F_t' + V_t \tag{2.19}$$

and $\tilde{\underset{\sim}{y}}^P_{1|0}$ and $B_{1|0}$ are set equal to $F_1 \, \tilde{\underset{\sim}{\theta}}^C_1$ and V_1 respectively. The above result is analogous to the equivalence of SSE from the usual least squares for LM and the sum of squares of one-step ahead prediction residuals obtained from the recursive least squares method.

Finally, it can be easily seen from (2.16) that if the model covariance matrices V_t and W_t are only specified up to a multiplicative overdispersion parameter σ^2, the estimates $\tilde{\underset{\sim}{\theta}}^P_1$ and $\tilde{\underset{\sim}{\theta}}^L_1$ do not change except for the multiplicative adjustment in their MSE by a factor of σ^2.

In the next section, we propose SSGLM - state space generalized linear models as an extension of SSLM. Notice that in the GLM extension to LM, the model was linearized by transforming from $\underset{\sim}{y}_t$ to $\underset{\sim}{z}_t$ via IWLS. Thus, it is natural to define SSGLM by applying SSLM on the transformed series $\{\underset{\sim}{z}_t\}$ i.e. the FWLS algorithm is administered on $\{\underset{\sim}{z}_t\}$. In other words, both filtering and iterative steps are required in order to obtain WLS estimates in SSGLM. This leads to the algorithm FIWLS for the proposed method. It may be noted that this algorithm is somewhat similar to the IWFLS (Iteratively Weighted and Filtered Least Squares) algorithm of Zeger (1988) which was introduced for a different purpose and does not use the recursive Kalman filter. In using FIWLS, we first need to specify the error covariance matrices $V_t(\underset{\sim}{\mu}_t(\underset{\sim}{\theta}_t))$ and W_t. For large n_t, the matrix V_t can be reasonally well approximated by $V_t(\underset{\sim}{\mu}_t(\hat{\underset{\sim}{\theta}}{}^C_t))$ where $\hat{\underset{\sim}{\theta}}{}^C_t$ is a consistent estimate of $\underset{\sim}{\theta}_t$ similar to the one given in (2.7). As regards W_t, if we can assume that it is time-invariant i.e. $W_t = W$, then for T large a consistent estimate can be constructed by using a method parallel to the one employed in RCR models of Swamy (1970). This is described in the following subsection.

2.4 Specification of the covariance matrix W_t under the time-invariance assumption

A consistent estimate $\hat{W}$ can be defined under the assumption $W_t = W$ when n_t and T both are large. In regression models with random coefficients proposed in econometrics for cross-sectional data, Swamy (1970) used least squares regression estimates $\tilde{\beta}_i$'s from several groups (or clusters) to estimate variance of the random regression component β_i; see also Pfeffermann and Nathan (1981). Although the problem of prediction in time series is quite different from the problem of estimating the underlying β (or some function of β_i's), the consistent cross-sectional estimates $\{\hat{\theta}^C_t, t=1, \ldots T\}$ can be used in a similar manner to estimate W. In Swamy's (1970) method, a bias corrected variance estimate is obtained. The unbiasedness property in our framework would correspond to asymptotic unbiasedness for n_t large. The estimate $\hat{W}$ can be defined as follows.

For $t \geq 2$, let

$$\hat{\beta}_t = \hat{\underset{\sim}{\theta}}{}^C_t - G_t\, \hat{\underset{\sim}{\theta}}{}^C_{t-1} \tag{2.20a}$$

$$R_1 = (T-1)^{-1} \textstyle\sum_{t=2}^{T} E[(\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)\ (\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)'] \tag{2.20b}$$

$$R_2 = (T-1)^{-1} \textstyle\sum_{t=2}^{T} E\ [\underset{\sim}{\beta}_t\ (\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)'], \tag{2.20c}$$

and define two estimators $\hat{W}$ and $\hat{\hat{W}}$ given by

$$\hat{W} = (T-1)^{-1} \textstyle\sum_{t=2}^{T} \hat{\underset{\sim}{\beta}}_t \hat{\underset{\sim}{\beta}}{}'_t, \quad \hat{\hat{W}} = \hat{W} - R_1 - R_2 - R_2'. \tag{2.20d}$$

We have

$$E(\hat{W}) = W + R_1 + R_2 + R_2', \quad E(\hat{\hat{W}}) = W \tag{2.21}$$

Therefore, bias in $\hat{W}$ is given by $R_1+R_2+R_2'$. Notice that the term R_2 is not zero because $E(\hat{\underset{\sim}{\beta}}_t | \underset{\sim}{\theta}_t, \underset{\sim}{\theta}_{t-1})$ is not, in general, equal to $\underset{\sim}{\beta}_t$. The bias corrected estimate $\hat{\hat{W}}$ of (2.20d) with suitable estimates of R_i's is analogous to Swamy's (1970) variance estimate. However, the bias term $R_1+R_2+R_2'$ would be negligible for large n_t when the mean and covariance of $\hat{\underset{\sim}{\beta}}_t$ conditional on $(\underset{\sim}{\theta}_t, \underset{\sim}{\theta}_{t-1})$ coincide in limit with those of the asymptotic distribution. The necessary regularity conditions for this to hold will be assumed in this article and therefore $\hat{W}$ would be (approximately) unbiased. Thus, for large n_t, we can omit bias correction and use only $\hat{W}$ to estimate W. It may be noted that the estimate $\hat{W}$, unlike $\hat{\hat{W}}$, is always non-negative definite which is, of course, desirable in practice. However unless T is large, $\hat{W}$ will not be consistent for W. It will be seen later in section 4 that this kind of misspecification of W when T is not large, does not affect the consistency of parameters estimates of the SSGLM predictor under fairly mild conditions.

3. THE PROPOSED METHOD — SSGLM

3.1 **Definition** The state space generalized linear model (SSGLM) can be defined in terms of the following two equations.

(i) Cross-sectional Behaviour: For each $t=1, \ldots, T$,

$$\underset{\sim}{y}_t = \underset{\sim}{\mu}_t + \underset{\sim}{\varepsilon}_t, \tag{3.1a}$$

$$\underset{\sim}{\eta}_t \equiv g(\underset{\sim}{\mu}_t) = F_t\, \underset{\sim}{\theta}_t, \tag{3.1b}$$

where $\underset{\sim}{\varepsilon}_t | \underset{\sim}{\mu}_t \sim WS(0, V_t(\underset{\sim}{\mu}_t))$, $Cov(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_s | \underset{\sim}{\mu}_t, \underset{\sim}{\mu}_s) = 0$ for $t \neq s$, and g is a monotone differentiable link function.

(ii) Longitudinal Behaviour: For $t=2, \ldots, T$,

$$\underset{\sim}{\theta}_t = G_t \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\xi}_t, \tag{3.2}$$

where $\underset{\sim}{\xi}_t \sim WS(0, W_t)$ along with the usual conditions as given earlier by (2.14).

The two main differences between this formulation and that of SSLM given by (2.11) and (2.13) are that the covariance matrix V_t depends on the mean vector $\underset{\sim}{\mu}_t$ and hence on the state vector $\underset{\sim}{\theta}_t$, and that the link function is not necessarily the identity. In fitting SSGLM to time series data, it will generally be assumed that both n_t and T are large. The choice of the design matrices F_t's can be guided by cross-sectional analyses and that of the G_t's by analysing the temporal pattern in the series of cross-sectional estimates $\{\hat{\underset{\sim}{\theta}}{}^C_t\}$. The covariance matrix W_t, if not known apriori, could be estimated by $\hat{W}$ under the assumption of time-invariance as described in the subsection 2.4. It should also be noted that the above formulation could be obviously extended to allow for the overdispersion parameter σ^2 as was the case with SSLM discussed earlier in subsection 2.3. For fitting SSGLM, we propose the following algorithm for estimation of model parameters.

3.2 Estimation Algorithm — FIWLS

The filtered and iterative weighted least squares (FIWLS) algorithm for estimating (or predicting) $\underset{\sim}{\theta}_T$ consists of two stages, each requring a series of iterative steps.

Stage I: Linearization for state space formulation

First transform $\underset{\sim}{y}_t$ to $\underset{\sim}{z}_t$ for each $t=1, \ldots, T$ as in (2.5). Now for n_t large, an approximate SSLM framework for $\{\underset{\sim}{z}_t\}$ series can be defined as

$$\underset{\sim}{z}_t = F_t \underset{\sim}{\theta}_t + \underset{\sim}{\delta}_t, \tag{3.3a}$$

$$\underset{\sim}{\theta}_t = G_t \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\xi}_t, \tag{3.3b}$$

where

$$\underset{\sim}{\delta}_t \doteq WS(0, U_t(\hat{\underset{\sim}{\theta}}{}^C_t)), \ \underset{\sim}{\xi}_t \sim WS(0, W_t), \tag{3.4a}$$

$$U_t(\hat{\underset{\sim}{\theta}}{}^C_t) = (d\underset{\sim}{\eta}_t/d\underset{\sim}{\mu}_t') \, V_t(\underset{\sim}{\mu}_t)(d\underset{\sim}{\eta}_t/d\underset{\sim}{\mu}_t')' \Big|_{\underset{\sim}{\theta}_t = \hat{\underset{\sim}{\theta}}{}^C_t} . \tag{3.4b}$$

The error vectors $\underset{\sim}{\delta}_t$, $\underset{\sim}{\xi}_t$ satisfy the usual conditions given earlier for the definition of SSLM in the subsection 2.3.

Stage II: Kalman Filtering for obtaining $\hat{\underset{\sim}{\theta}}{}^L_T$

The BLUP $\hat{\underset{\sim}{\theta}}{}^L_T$ (only approximate in view of the linearization in stage I) of $\underset{\sim}{\theta}_T$ based on $\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_T$ can be computed in the same way as $\tilde{\underset{\sim}{\theta}}{}^L_T$ was obtained from the KF given in (2.16). The appropriate modifications of (2.16) are obtained by replacing $\underset{\sim}{\theta}_t$, V_t and $A_{t|t-1}$ by $\hat{\underset{\sim}{\theta}}_t$, U_t, and $C_{t|t-1}$ respectively. The KF is started off by $\hat{\underset{\sim}{\theta}}{}^L_1$ and C_1 where $\hat{\underset{\sim}{\theta}}{}^L_1$ is the cross-sectional WLS estimate $\hat{\underset{\sim}{\theta}}{}^C_1$ as in (2.7) for fixed $\underset{\sim}{\theta}_1$ when U_1 is substituted for Σ_1, and C_1 is $(F_1' U_1^{-1} F_1)^{-1}$. As mentioned earlier in the introduction the GKF (Generalized Kalman Filter) proposed by Zehnwirth (1988) for state-dependent observation variance and identity link function is related to the above KF for the model defined by (3.3) and (3.4) in the sense that Zehnwirth uses $\bar{U}(= E_{\underset{\sim}{\theta}} \, U_t(\underset{\sim}{\theta}_t))$ and not $U_t(\hat{\underset{\sim}{\theta}}{}^C_t)$ in defining KF. This essentially amounts to approximating $\bar{U}$ by the expression inside the expection. $\bar{U}$ would generally be computationally intractable for non-linear link function g. However, if $\bar{U}$ were available, it should be preferable in the interest of optimality. We can also calculate the error sum of squares for the model (3.3), analogous to the expression (2.18), as a by-product of Kalman filtering as follows:

$$SSE = \sum_{t=1}^{T} (\underset{\sim}{z}_t - \hat{\underset{\sim}{z}}{}^P_{t|t-1})' \, D^{-1}_{t|t-1} \, (\underset{\sim}{z}_t - \hat{\underset{\sim}{z}}{}^P_{t|t-1}), \tag{3.5}$$

where $\hat{z}^P_{1|0}$ and $D_{1|0}$ are defined as $F_1 \hat{\underset{\sim}{\theta}}{}^C_1$ and $U_1(\hat{\underset{\sim}{\theta}}{}^C_1)$ respectively and $D_{t|t-1}$ as in (2.18) with A replaced by C and V by U. From SSE, an estimate of the overdispersion parameter σ^2 can be obtained as SSE/DF where DF denotes the appropriate degrees of freedom. After fitting the model, we next consider some methods for model checking.

3.3 Diagnostics

Let the data be available up to time $T+T'$. Suppose data for the first T (chosen arbit-rarily) points are used for model fitting. We shall refer to diagnostics based on these points as "within sample" and those based on time points $T+1, \ldots, T+T'$ as "post sample". The following tools can be used for checking fit of the model, see e.g. Harvey (1984), Harvey and Durbin (1986), and Harvey and Fernandes (1989).

3.3.1 Within Sample Diagnostics

(a) Let r_{it} denote the standardized one-step ahead prediction residual corresponding to the ith element of vector $\underset{\sim}{z}_t$ at time t. These residuals for each i, can be plotted against time and against $\tilde{z}^P_{it|t-1}$ and examined for randomness.

(b) Check whether the sample variance of the residuals $\{r_{it}:\ t=2, \ldots, T\}$ for each i is close to one. A value greater than one indicates overdispersion relative to the model being fitted (Harvey and Fernandes, 1989).

(c) Following Harvey (1984), first a naive model is chosen as a yardstick which is defined by

$$g(\underset{\sim}{y}_t) = g(\underset{\sim}{y}_{t-1}) + \underset{\sim}{\beta} + \underset{\sim}{\zeta}_t,\ \underset{\sim}{\zeta}_t \sim WS(0, \sigma^2 I), \tag{3.6}$$

where g is the link function defined by (3.1b) and $\underset{\sim}{\beta}$ is a constant drift parameter. Next, the root mean square of the one-step ahead prediction errors within sample for grouped data case is computed as

$$RMSEW_0 = [\textstyle\sum_{t=2}^{T}(\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}^P_{t|t-1})'(\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}^P_{t|t-1})/(m(T-1)-k_0)]^{\frac{1}{2}} \tag{3.7}$$

where $\hat{y}^P_{-t|t-1}$ is $g^{-1}(g(\underset{\sim}{y}_{t-1}) + \hat{\underset{\sim}{\beta}})$, $\hat{\underset{\sim}{\beta}}$ denotes the average of the first differences of $g(\underset{\sim}{y}_t)$'s, and k_0 denotes the length of $\underset{\sim}{\beta}$ in the model (3.6). Similarly, for the model of interest defined by (3.3) and (3.4), we compute

$$RMSEW_1 = [\textstyle\sum_{t=2}^{T}(\underset{\sim}{y}_t - \hat{\underset{\sim}{y}}^P_{t|t-1})'(\underset{\sim}{y}_t - \hat{\underset{\sim}{y}}^P_{t|t-1})/(m(T-1) - k_1)]^{\frac{1}{2}}, \tag{3.8}$$

where k_1 is the number of fixed parameters estimated in order to apply the linear predictor. For the ungrouped data case, the denominators in (3.7) and (3.8) are suitably modified. If $RMSEW_1$ is more than $RMSEW_0$, then clearly the model is not worth pursuing.

3.3.2 Post-sample Diagnostics

(a) Post-sample predictive tests for the grouped data case can be defined. See the next section for their asymptotic justification when n_t and T are large. With τ-step ahead predictions $\hat{z}^P_{T+\tau|T}$, defined below in the subsection 3.4, a chi-square test for lack of fit of the model is given by rejecting for large values of X^2_τ (referred to a χ^2_m distribution) where for $\tau=1, 2, \ldots, T'$,

$$X^2_\tau = (\underset{\sim}{z}_{T+\tau} - \hat{\underset{\sim}{z}}^P_{T+\tau|T})'\ D^{-1}_{T+\tau|T}\ (\underset{\sim}{z}_{T+\tau} - \hat{\underset{\sim}{z}}^P_{T+\tau|T}). \tag{3.9}$$

(b) Cross-validation errors can be computed for both the naive model and the model of interest and examined for the extent of improvement. For τ-step ahead predictions, cross-validation errors for the grouped data case can be defined by the root mean square of τ-step ahead prediction errors for the post-sample as

$$RMSEP_0(\tau) = [\sum_{j=0}^{T'-\tau}(\underset{\sim}{y}_{T+\tau+j} - \tilde{\underset{\sim}{y}}^P_{T+\tau+j|T+j})'(\underset{\sim}{y}_{T+\tau+j} - \tilde{\underset{\sim}{y}}^P_{T+\tau+j|T+j})/m(T'-\tau+1)]^{\frac{1}{2}} \tag{3.10}$$

where $\tilde{\underset{\sim}{y}}^P_{T+\tau+j|T+j}$ are predictions for the naive model. Similarly, $RMSEP_1(\tau)$ for the model of interest can be defined using the untransformed vectors $\underset{\sim}{y}_t$ and their predictors. The denominator in (3.10) in the case of ungrouped data can be modified in an obvious manner.

3.4 Prediction and Smoothing

Predictions for the post sample period are needed for diagnostic purposes and if the model is deemed adequate, then predictions for future observations and their associated MSE's (mean square error) would generally be required. For this purpose, the updating equations of the KF are simply bypassed, and the BLUP of $\underset{\sim}{\theta}$, τ periods ahead, is first obtained recursively as

$$\hat{\underset{\sim}{\theta}}^P_{T+\tau|T} = G_{T+\tau}\ \hat{\underset{\sim}{\theta}}^P_{T+\tau-1|T}, \tag{3.11}$$

and its MSE as
$$C_{T+\tau|T} = G_{T+\tau}\, C_{T+\tau-1|T}\, G'_{T+\tau} + W_{T+\tau} \tag{3.12}$$

Note that for predicting at $t>T+T'$, all the data up to $T+T'$ time points should be used by refitting the model. Now, the predictor of $\underset{\sim}{z}_{T+\tau}$ is obtained as
$$\underset{\sim}{\hat{z}}^{P}_{T+\tau|T} = F_{T+\tau}\, \underset{\sim}{\hat{\theta}}^{P}_{T+\tau|T}\ , \tag{3.13}$$
and the corresponding MSE is
$$D_{T+\tau|T} = F_{T+\tau}\, C_{T+\tau|T}\, F'_{T+\tau} + U_{T+\tau}\ . \tag{3.14}$$

Here $U_{T+\tau}$ can be evaluated at $\underset{\sim}{\hat{\theta}}^{P}_{T+\tau|T}$, the τ-step ahead predictor of $\theta_{T+\tau}$ whenever $\underset{\sim}{\hat{\theta}}^{C}_{T+\tau}$ is not available. For the untransformed $\underset{\sim}{y}_t$, the $\hat{y}^{P}_{\sim T+\tau|T}$ is obtained as $g^{-1}(\underset{\sim}{\hat{z}}^{P}_{T+\tau|T})$ and the MSE for $\underset{\sim}{y}\ -\hat{y}^{P}_{\sim T+\tau|T}$ is approximately given by $(d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}'_t)\ D_{T+\tau|T}\ (d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}'_t)'$.

The smoothed predictors of $\underset{\sim}{\theta}_t$ or $\underset{\sim}{y}_t$ for any point $t<T$ given all the observations $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_T$ can be carried out using the algorithm given in Harvey (1981, p. 115) or a fast algorithm due to Kohn and Ansley (1989).

4. SOME THEORETICAL RESULTS

Suppose n_t is large for each t so that $U_t(\underset{\sim}{\hat{\theta}}^{C}_t)$'s do provide approximate covariance matrices for $\underset{\sim}{z}_t$'s conditional on θ_t's. The covariance matrix W_t is assumed known for propositions 4.1 and 4.2. In proposition 4.3, however, we investigate the effect of misspecified W on the estimates of $\underset{\sim}{\theta}_t$ when T is not large i.e. when $\hat{W}$ is unstable. The following proposition establishes the asymptotic distribution of X^2_τ defined earlier by (3.10).

<u>Proposition 4.1</u> Suppose the data are grouped into m-vectors $\underset{\sim}{y}_t$'s which for large n_t, are asymptotically normal under a suitable CLT. Then
$$X^2_\tau \doteq \chi^2_m \tag{4.1}$$
To see this, observe that
$$\underset{\sim}{z}_t - F_t\, \underset{\sim}{\theta}_t \doteq N_m(0,\ U_t(\underset{\sim}{\hat{\theta}}^{C}_t)),$$
$$\underset{\sim}{\hat{z}}^{P}_{t|t-1} - F_t\, \underset{\sim}{\theta}_t \doteq N_m(0,\ F_t\, C_{t|t-1}\, F'_t)$$
Thus under our model assumptions, since $\underset{\sim}{z}_t$ and $\underset{\sim}{\hat{z}}^{P}_{t|t-1}$ are uncorrelated given $\underset{\sim}{\theta}_t$, we get
$$\underset{\sim}{z}_t - \underset{\sim}{\hat{z}}^{P}_{t|t-1} \doteq N_m(0,\ D_{t|t-1}),$$
where $D_{t|t-1}$ is $F_t\, C_{t|t-1}\, F'_t + U_t(\underset{\sim}{\hat{\theta}}^{C}_t)$. The result (4.1) then follows immediately.

Now consider the grouped data situation in which the measurement equation represents a saturated model for $\underset{\sim}{\mu}_t$, i.e. the vector $\underset{\sim}{\theta}_t$ contains m-parameters. Cross-sectionally, the optimal predictor of $\underset{\sim}{y}_t$ is $\underset{\sim}{y}_t$ itself. Longitudinally, one could also show that by allowing for nonrandom drift parameters in the transition equation, $\underset{\sim}{\hat{y}}^{L}_t$ is $\underset{\sim}{y}_t$ itself. Also $\underset{\sim}{\hat{z}}^{P}_{t|t-1}$ (or $\underset{\sim}{\tilde{y}}^{P}_{t|t-1}$) is equal to $\underset{\sim}{z}_t$(or $\underset{\sim}{y}_t$) which implies that SSE is zero. This is given by the next proposition.

<u>Proposition 4.2</u> For the grouped case, let $\underset{\sim}{\theta}_t$ represent parameters in the saturated model for cross-sectional behaviour and $\underset{\sim}{\gamma}_t$ represent unknown but nonrandom time varying drift parameters in modelling the longitudinal behaviour i.e. the transition equation is given by, for $t\geq 2$,
$$\underset{\sim}{\theta}_t = G_t\, \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\gamma}_t + \underset{\sim}{\zeta}_t\,. \tag{4.2}$$
Then
$$\underset{\sim}{\hat{y}}^{L}_t = \underset{\sim}{y}_t\ , \tag{4.3a}$$
and
$$\underset{\sim}{\hat{y}}^{P}_{t|t-1}(\underset{\sim}{\hat{\gamma}}_t) = \underset{\sim}{y}_t. \tag{4.3b}$$

To prove this, notice that the reduced form (2.14) will now be modified to contain parameters $\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_T$ along with $\boldsymbol{\theta}_T$. Then the number of parameters is the same as the number of z_{it}'s, i.e. mT. Hence the result (4.3a) follows because $\boldsymbol{y}_t$ is $F_t^{-1}(\boldsymbol{z}_t)$ for the saturated model. To derive (4.3b), first note that for given $\boldsymbol{\gamma}_t$'s, $\hat{\boldsymbol{z}}^P_{t|t-1}$ is $F_t\ G_t\ \hat{\boldsymbol{\theta}}^L_{t-1} + F_t\ \boldsymbol{\gamma}_t$. By substituting estimates of $\boldsymbol{\gamma}_t$ obtained by minimizing SSE of (3.5) up to time t, it can be seen that $\hat{\boldsymbol{z}}^P_{t|t-1}$ is $\boldsymbol{z}_t$ and hence (4.3b) follows.

Next consider the situation when the unknown $W_t = W$ but T is not large. In this case $\hat{W}$ will not be a good specification for W because it is unlikely to be in the proximity of W. Although the BLUP property of estimates will no longer hold, the next proposition shows that the estimates $\hat{\boldsymbol{\theta}}^P_T$ and $\hat{\boldsymbol{\theta}}^L_T$ continue to be consistent for large n_t. This property of robust inference about $\boldsymbol{\theta}_t$ is similar to the one obtained by Zeger (1988) for regression models for time series of counts.

<u>Proposition 4.3</u> Assume that the mean function is correctly specified in terms of F_t and G_t, but W may be misspecified by $\hat{W}$ when T is not large. Then, for large n_t and for $t \geq 2$, the asymptotic means of the distributions of $\hat{\boldsymbol{\theta}}^P_{t|t-1} - \boldsymbol{\theta}_t$ and $\hat{\boldsymbol{\theta}}^L_{t|t} - \boldsymbol{\theta}_t$ remain the same, i.e. zero, but their MSEs change to $C^{\star}_{t|t-1}$ and $C^{\star}_t$ respectively where, for $t \geq 2$,

$$C^{\star}_{t|t-1} = G_t\ C^{\star}_{t-1}\ G_t' + W, \tag{4.4}$$

$$C^{\star}_t = (I - \hat{K}_t F_t)\ \hat{C}_{t|t-1}(\hat{C}^{-1}_{t|t-1}\ C^{\star}_{t|t-1}\ \hat{C}^{-1}_{t|t-1} + F_t'\ U_t^{-1} F_t)\ \hat{C}_{t|t-1}(I - \hat{K}_t F_t)'\ , \tag{4.5}$$

and "^" indicates that $\hat{W}$ is substituted for W. In (4.4) above, for $t=1$, $C^{\star}_1$ is the same as C_1 defined earlier by $(F_1'\ U_1^{-1}\ F_1)^{-1}$.

The proof of the above proposition can be seen as follows. Following Zehnwirth (1988), the estimator $\hat{\boldsymbol{\theta}}^L_t$ can be expressed as a linear combination of $\hat{\boldsymbol{\theta}}^P_{t|t-1}$ and the estimator $\hat{\boldsymbol{\theta}}^C_t$ or $(F_t' U_t^{-1}\ F_t)^{-1}\ F_t'\ U_t^{-1}\ \boldsymbol{z}_t$, i.e.

$$\hat{\boldsymbol{\theta}}^L_t = (I - \hat{\Lambda})\ \hat{\boldsymbol{\theta}}^P_{t|t-1} + \hat{\Lambda}\ \hat{\boldsymbol{\theta}}^C_t \tag{4.6}$$

where

$$\Lambda = (I - K_t F_t)\ C_{t|t-1}\ F_t'\ U_t^{-1}\ F_t, \tag{4.7}$$

and $\hat{\Lambda}$ corresponds to Λ when $\hat{W}$ is substituted for W. As $n_t \to \infty$ the consistency of $\hat{\boldsymbol{\theta}}^L_t$ now follows easily by induction starting with $t=2, 3, \ldots$ and so on. To obtain the expression (4.4) for $C^{\star}_t$, write another equivalent expression for (4.6) as

$$\hat{\boldsymbol{\theta}}^L_t = \hat{\boldsymbol{\theta}}^P_{t|t-1} + \hat{K}_t(\boldsymbol{z}_t - F_t\ \hat{\boldsymbol{\theta}}^P_{t|t-1}) \tag{4.8a}$$

$$= (I - \hat{K}_t F_t)\ \hat{C}_{t|t-1}(\hat{C}^{-1}_{t|t-1}\ \hat{\boldsymbol{\theta}}^P_{t|t-1} + F_t'\ U_t^{-1}\ \boldsymbol{z}_t), \tag{4.8b}$$

because of the identity

$$(I - K_t F_t)\ C_{t|t-1}\ F_t' U_t^{-1} = K_t\ . \tag{4.9}$$

The desired results (4.4) and (4.5) follow immediately from (4.8b). It may be noted that when $\hat{W} = W$, then in view of (4.9), $C^{\star}_t$ of (4.5) reduces approximately to C_t or $(I - K_t F_t)\ C_{t|t-1}$ as expected.

5. APPLICATION TO CANCER MORTALITY SERIES OF COUNTS

For the purpose of illustrating SSGLM, the data on annual lung cancer mortality counts for Ontario for the years 1970-1987 cross-classified by sex and age was analysed. Five age groups were considered: $1 \equiv$ 0-44, $2 \equiv$ 45-54, $3 \equiv$ 55-64, $4 \equiv$ 65-74 and $5 \equiv$ 75+. The ten time series of counts are shown in Figures 1 and 2 classified by male and female. We used data for 1970-85 (i.e. 16 time points) to fit the model and the data for the next two time points (1986, 1987) for post-sample diagnostics.

For fitting SSGLM, we first need to specify the cross-sectional behaviour at time t. There are ten groups i.e. $m=10$. For the ith group, the count y_{it}, for each i is assumed to follow a Poisson-motivated wide sense distribution with mean μ_{it} equal to $n_{it}\ \lambda_{it}$ where n_{it} is the known population size. A log-linear model for the rates $\{\lambda_{it},\ i=1,\ \dots,\ 10\}$ was considered. A row-effects model for ordinal data where row and column refer to sex and age respectively (Agresti, 1984, p. 84) gave a reasonable fit cross-sectionally for almost all time points. Therefore, the cross-sectional model with suitable scores for age categories was chosen as

$$\log \lambda_{it} = \mathbf{f}_i' \boldsymbol{\theta}_t,\ i=1,\ 2,\ \dots,\ 10 \tag{5.1a}$$

where

$$F = \begin{bmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{f}_{10}' \end{bmatrix} = \left[\begin{array}{ccccccc} 1 & 1 & 1 & 0 & 0 & 0 & -2 \\ 1 & 1 & 0 & 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 2 \\ \hdashline 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \tag{5.1b}$$

There is a total of seven effects $\theta_1,\ \theta_2, \dots, \theta_7$, one for constant, one for sex, four for age, and one for age-sex interaction. Note that the incidence matrix F_t is time-invariant in this case. The covariance matrix $U_t(\boldsymbol{\theta}_t)$ is evaluated at the estimate $\hat{\boldsymbol{\theta}}_t^C$ as

$$U_t(\hat{\boldsymbol{\theta}}_t^C) = \operatorname{diag}\ (\hat{\boldsymbol{\theta}}_t^C)^{-1} \tag{5.2}$$

The form of (5.2) is easily obtained using the Poisson variance-mean relation and the log link function.

Next for specifying the longitudinal behaviour in SSGLM, the plots of $\hat{\theta}_{it}^C$ and $\hat{\theta}_{it}^C - \hat{\theta}_{it-1}^C$ against time were examined for each $i=1,\ \dots,\ 10$; see figures 3 and 4. The first difference series of $\hat{\theta}_{it}^C$ appeared fairly random around zero mean except for a slight drift in $\hat{\theta}_{1t}^C$ series. One could regress $\hat{\theta}_{it}^C$ on $\hat{\theta}_{it-1}^C$ for each i and check for randomness in the residuals instead of the first differences. However, for the sake of illustration and simplicity, a random walk with no drift model was chosen to represent the transition equation, i.e.

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\xi}_t \tag{5.3}$$

Thus, the transition matrix G_t was also assumed to be time invariant and set equal to I. The covariance of $\boldsymbol{\xi}_t$ was estimated by

$$\hat{W} = \frac{1}{T-1} \sum_{t=2}^{T} (\hat{\boldsymbol{\theta}}_t^C - \hat{\boldsymbol{\theta}}_{t-1}^C)(\hat{\boldsymbol{\theta}}_t^C - \hat{\boldsymbol{\theta}}_{t-1}^C)' \tag{5.4}$$

After having specified SSGLM, the model was fitted using the FIWLS algorithm given earlier in subsection 3.2. The sample standard deviations of standardized one-step ahead residuals r_{it}'s, $i=1,\ \dots,\ 10$ are given in Table 1.

Table 1: Estimated Standard Deviations of Standardized One-step Ahead Prediction Residuals

i	1	2	3	4	5	6	7	8	9	10
SD(r_{it})	.76	.86	.70	.99	.50	.88	.69	.68	.99	.72

There is no indication of overdispersion because all $\mathrm{SD}(r_{it})$'s are below one. A plot of r_{it}'s against time showed no indication of misspecification. Post-sample predictive tests for one-step and two-step ahead projections were carried out for the last two years 1986 and 1987. The X^2 values were obtained as

$$X_1^2 = 9.40 \quad , \quad X_2^2 = 1.14\ , \tag{5.5}$$

which when referred to a χ_{10}^2 distribution were clearly insignificant. The SSE was computed as 167.5 with an estimate of the overdispersion parameter σ as 1.046. Again there seemed no indication of overdispersion. The RMSEW_1, $\mathrm{RMSEP}_1(1)$ and $\mathrm{RMSEP}_1(2)$ were obtained respectively as 31.1, 28.1, and 17.6. The corresponding values for the naive model were 84.7, 26.4 and 56.1. The SSGLM seems to give a considerable improvement over the naive model. It may be noted that in computing RMSEP(1), only predictions for one time point (1986) were made using data up to 1985.

Figures 5 and 6 show sample plots of actual, cross-sectionally fitted, one-step ahead predicted (two-step for the last point), and filtered (or updated) counts for male in the age group 65-74 and female in the age group 55-64 respectively. Table 2 gives a summary of predicted counts as well as actual counts over all the 10 groups for 1986 and 1987. The values of RMSE for predicted counts are given in the parentheses. These are not likely to be stable because T is not large in the example under consideration.

Table 2: Predicted counts (P) vs. Actual counts (A) for Lung Cancer Mortality in Ontario

		Male					Female				
Age Group:		1	2	3	4	5	1	2	3	4	5
1986	P:	58 (14)	246 (36)	818 (64)	1114 (74)	799 (57)	46 (14)	143 (20)	370 (42)	424 (53)	317 (55)
	A:	51	242	851	1050	775	38	150	329	435	304
1987	P:	59 (18)	250 (48)	823 (83)	1155 (99)	833 (74)	47 (18)	146 (25)	371 (55)	441 (74)	329 (76)
	A:	56	256	810	1110	835	50	164	383	422	334

6. DISCUSSION

It is shown that if the number of time points and the number of observations at each time point are fairly large, then a nonstationary and non-normal time series data under possibly non-linear models can be suitably transformed for application of state space linear modelling techniques. The consistent cross-sectional parameter estimates $\{\hat{\underset{\sim}{\theta}}{}_t^C\}$ can be used to specify approximately the covariance matrix W of the transition equation when T is large and W_t is assumed time-invariant. It may be noted that if the transition matrix G_t involves some unknown parameters, they can also be estimated consistently by using Zellner's (1962) two-step Aitken estimator introduced in the context of seemingly unrelated regression equations. It is also shown that when T is not large, inferences about $\underset{\sim}{\theta}_t$ remain robust to misspecification of W provided the mean function is correctly specified.

As the Kalman filter approach (Harvey, 1984) can routinely handle missing data problems when observations are assumed to be equi-spaced, the proposed method SSGLM can also be applied to these situations. There are certain directions, however, in which extensions of SSGLM could be investigated. For instance, inclusion of seasonal effects for monthly or quarterly series as well as intervention effects in the transition equation for SSGLM would be desirable. The present SSGLM framework can be modified to include nonstochastic seasonal or intervention effects. However, the case of stochastic effects needs further investigation. Also for time series arising from complex surveys, it would be important to investigate the impact of complex designs on the inference about model parameters analogous to Rao and Scott (1984) adjustments for cross-sectional data analysis. In the case of panel surveys, there is the additional problem of correlated survey errors in the measurement equation due to overlapping units between successive time points as considered by Binder and Dick (1989) and Pfeffermann (1989) for ARMA modelling of survey errors in the context of state space linear models.

ACKNOWLEDGEMENTS

The first author's research was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University as an adjunct research professor. Thanks are due to D. Binder and J.N.K. Rao for their useful comments and especially D. Pfeffermann for his insightful remarks and several helpful discussions. Thanks are also due to Christine Larabie, Carmen Lacroix, and especially Judy Clarke for the processing of the manuscript in a very short time.

REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: John Wiley.

Binder, D.A. and Dick, P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology, 15,* 29-45.

Box, G.E.P. and Jenkins, G.M. (1970). *Time series analysis: forecasting and control.* San Francisco: Holden-Day.

Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics, 42,* 693-734.

Cox, D.R. (1981). Statistical analysis of time series, some recent developments (with discussion). *Scand. Jour. Statist., 8,* 93-115.

Fuller, W.A. (1976). *Introduction to Statistical Time Series.* New York: John Wiley.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics, 25*, 489-504.

Harvey, A.C. (1981). *Time Series Models*, Oxford: Philip Allan, and New York: John Wiley.

Harvey, A.C. (1984). A unified view of statistical forecasting procedures (with discussion). *Jour. Forecasting, 3*, 245-275.

Harvey, A.C. and Peters, S. (1984). Estimation procedures for structural time series models. *London School of Economics, Discussion Paper* A. 44.

Harvey, A.C. and Durbin, J. (1986). The effects of seat belt legislation on British road casualties: A case study in statistical time series modelling (with discussion). *Jour. Roy. Statist. Soc. A, 149*, 187-227.

Harvey, A.C. and Fernandes, C. (1989). Time series models for counts or qualitative observations. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Kalbfleisch, J.D. and Lawless, J.F. (1984). Least-squares estimation of transition probabilities from aggregate data. *Can. Jour. Statist. 12*, 169-182.

Kalbfleisch, J.D. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Jour. Amer. Statist. Assoc., 80*, 863-871.

Kaufmann, H. (1987). Regression methods for non-stationary categorical time series: Asymptotic estimation theory. *Ann. Statist., 17*, 79-98.

Kitagawa, G. (1987). Non-Gaussian state space modelling for non-stationary time series (with discussion). *Jour. Amer. Statist. Assoc., 82*, 1032-1063.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics, 33*, 133-158.

Kohn, R. and Ansley, C.F. (1989). A fast algorithm for signal extraction, influence, and cross-validation in state space models. *Biometrika*, 65-79.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika, 74*, 247-258.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist., 11*, 59-67.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* (2nd ed.). London: Chapman and Hall.

Pfeffermann, D. (1989). Estimation and seasonal adjustment of population means using data from repeated surveys. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from a cluster sample. *Jour. Amer. Statist. Assoc., 76*, 681-689.

Preisler, H. (1989). Analysis of a toxicological experiment using a generalized linear model with nested random effects. *Int. Statist. Rev. 57*, 145-159.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Ann. Statist., 12*, 46-60.

Smith, T.M.F. and Brunsdon, T.M. (1989). The time series analysis of compositional data. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Stiratelli, R., Laird, N.M., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40*, 961-972.

Stram, D.O., Wei, L.J., and Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Jour. Amer. Statist. Assoc., 83*, 631-637.

Swamy, P.A.V.B. (1970). Efficient inferences in a random coefficient regression model. *Econometrica, 38*, 311-323.

West, M., Harrison, J.P. and Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *Jour. Amer. Statist. Assoc., 80*, 73-96.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika, 75*, 621-629.

Zeger, S.L., Liang, K.-Y. and Self, S.G. (1985). The analysis of binary longitudinal data with time independent covariates. *Biometrika, 72*, 31-38.

Zeger, S.L., and Qaqish, B. (1988). Markov regression models for time series; a quasi-likelihood approach. *Biometrics*, 44,1019-1031.

Zehnwirth, B. (1988). A generalization of the Kalman filter for models with state-dependent observation variance. *Jour. Amer. Statist. Assoc., 83*, 164-167.

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregation bias. *Jour. Amer. Statist. Assoc.* 57, 348-368.

LUNG CANCER MORTALITY
MALE COUNTS

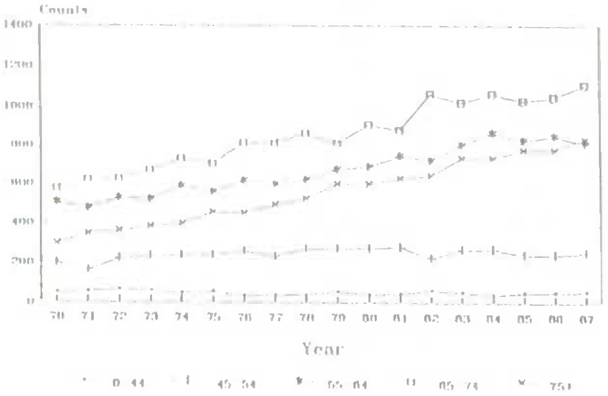


Figure 1

LUNG CANCER MORTALITY
FEMALE COUNTS

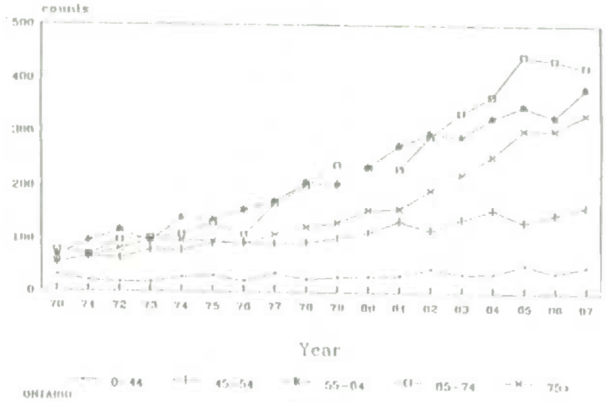


Figure 2

Cross-sectional theta values
1970 to 1987

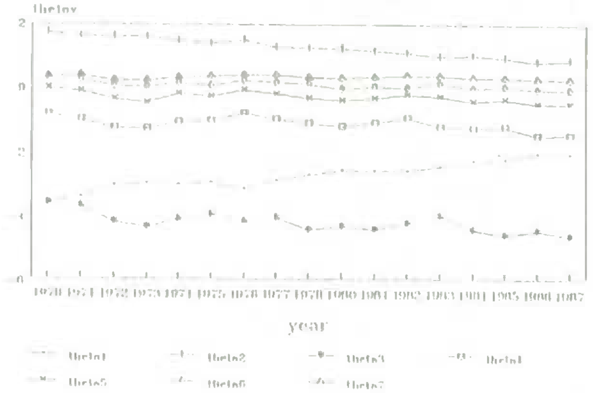


Figure 3

First difference theta values
1971-1987

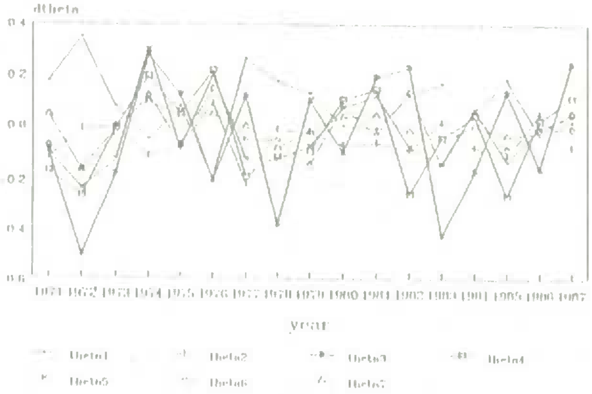


Figure 4

LUNG CANCER MORTALITY COUNTS
MALE AGE 65-74 NO DRIFT

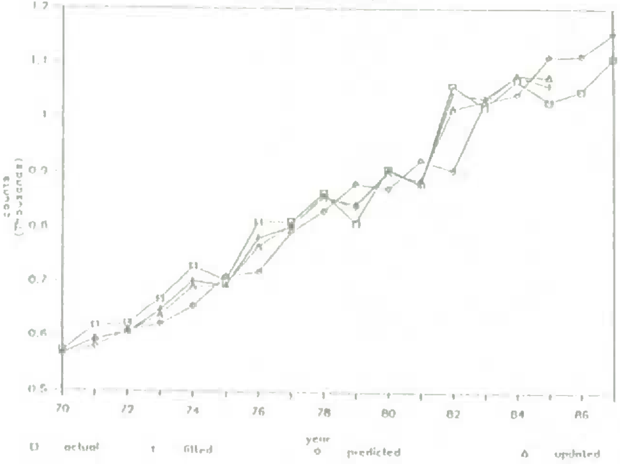


Figure 5

LUNG CANCER MORTALITY COUNTS
FEMALE AGE 55-64 NO DRIFT

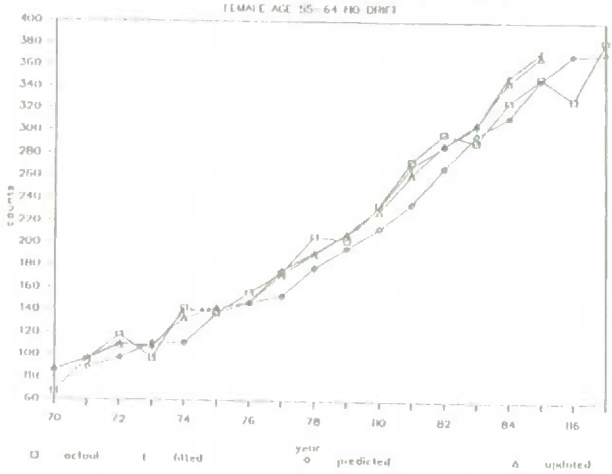


Figure 6

PART 4

DEVELOPMENTS IN THE ANALYSIS OF TIME SERIES DATA

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

ALTERNATIVE APPROACHES TO THE ANALYSIS OF TIME SERIES COMPONENTS

W. R. Bell and M. G. Pugh[1]

ABSTRACT

In the time series literature of recent years one finds different approaches to the analysis of time series postulated to follow some type of component structure. There are alternatives to the now familiar ARIMA (autoregressive– integrated–moving average) modeling approach, perhaps the most popular being the "structural modeling" approach of Harvey and others, which uses an explicit components structure. Despite the considerable research on these models, remarkably little work has appeared comparing results from the alternative approaches. Questions arise regarding the comparative fit of alternative models, and the effect of model choice on applications such as model–based seasonal adjustment and use of time series methods in repeated survey estimation. As these are empirical questions, we attempt to address them here through comparing results from applying such alternative models to some Census Bureau time series.

KEY WORDS: ARIMA Model; Components Model; AIC; Seasonal Adjustment; Repeated Survey Estimation.

1. INTRODUCTION

The analysis of the components of time series has a long history (discussed in Nerlove, Grether, and Carvalho 1979), going back to work in astronomy, meteorology, and economics in the 17th through 19th centuries, and to early seasonal analysis by Buys–Ballot (1847). Empirical methods of seasonal adjustment were developed in the early part of this century leading utlimately to the development of the well–known X–11 method in 1967. As discussed in Bell and Hillmer (1984), these methods were developed in advance of adequate seasonal time series models, which have only become widely available and computationally feasible in the last 20 years or so.

This well–established interest in time series components has had important influences on time series modeling; in particular, it has led to two rather different approaches to modeling and model–based seasonal adjustment. For the autoregressive–integrated–moving average (ARIMA) models (Box and Jenkins 1976), several approaches to seasonal adjustment have been developed. The most successful of these, in our view, is the "canonical" approach of Burman (1980) and Hillmer and Tiao (1982). In contrast, a "component modeling" approach has developed that uses simple ARIMA models for seasonal, trend, irregular, etc. components. This approach is exemplified in the work of Akaike (1980), Gersch and Kitagawa (1983) and Kitagawa and Gersch (1984), and Harvey and Todd (1983) and Harvey (1985). Nerlove, Grether, and Carvalho (1979) suggested a somewhat different approach that appears not to have caught on, possibly because their ARIMA component models are too flexible to even assure that the model structure is identified (Hotta 1989), and because their treatment of nonstationarity (by polynomial detrending) is now viewed as inadequate.

While there has been considerable developmental work on both modeling appproaches, there is surprisingly little literature comparing results for the two different approaches. Harvey and Todd (1983) compared the forecast performance of their "basic structural model" (BSM) with that of ARIMA models fitted by Prothero and Wallis (1976) to six quarterly macroeconomic time series. Their results were rather inconclusive, also some of the ARIMA models used were of unusual form, featuring long lags in the seasonal operators. (In fairness, Prothero and Wallis' (1976) work was in the early stages of development of seasonal ARIMA modeling, before such refinements as exact maximum likelihood and outlier treatment were readily available.) Expanding the BSM, Harvey (1985) developed components models to explain cyclical behavior (with nonseasonal series) and gave some discussion of their relation to ARIMA models. Maravall (1985) observed that the BSM could yield an overall model close to Box and Jenkins (1976) popular ARIMA $(0,1,1)\times(0,1,1)_{12}$ "airline model," by showing that autocorrelations for the differenced series could be similar for the two models (depending on parameter values). This raised the important possibility that the BSM and certain ARIMA models could be about the same for some series. Carlin and Dempster (1989), in a detailed analysis of two series, found only small differences between canonical ARIMA seasonal adjustments and those from a fractionally–integrated–moving average (FRIMA) components model, and more major differences when comparing the FRIMA adjustment with the X–11 adjustment used in practice for another series.

[1] W. R. Bell, Statistical Research Division, U. S. Bureau of the Census, Washington, D.C. 20233, U.S.A., M. G. Pugh, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

The literature seems to leave two important questions unanswered, namely: (1) do ARIMA or components models provide a better fit to actual data or can available data even discriminate between them, and (2) how different are the results from ARIMA and components models in practical applications? The former question is one of statistical significance, the latter one of practical significance. Both questions are largely empirical, and an empirical investigation into them shall be the primary focus of this paper. In section 2 we describe the specific models we shall consider in detail, and use the AIC criterion of Akaike (1973) to compare the fit of ARIMA models and the BSM for a set of 45 seasonal time series. In general, AIC expresses a strong preference for ARIMA models.

Section 3 considers seasonal adjustment. Bell and Hillmer (1984) noted that component modelers have ignored the inherent uncertainty about seasonal–nonseasonal decompositions consistent with any given fitted model. To address this we consider the range of admissible decompositions consistent with a given components model, and present a "canonical decomposition" for component models analogous to that given for ARIMA models by Burman (1980) and Hillmer and Tiao (1982). The canonical decomposition turns out to be trivially simple to obtain and very easy to use in signal extraction for seasonal adjustment. However, it also turns out to be very close to the original fitted components model for the series considered here, suggesting that seasonal adjustments for the original and canonical components models may typically be virtually identical. We then compare ARIMA model and BSM seasonal adjustments for two series and find negligible differences in signal extraction point estimates and proportionally large differences in signal extraction variances, though the signal extraction variances all seem small in an absolute sense.

In section 4 we investigate the effects of using ARIMA versus component models in applying time series signal extraction techniques to estimation for repeated surveys. This idea was originally suggested by Scott and Smith (1974) and Scott, Smith, and Jones (1977), but has seen intensive investigation more recently following theoretical and computational developments in estimation and signal extraction for nonstationary time series models. For the two series we consider the signal extraction point estimates using ARIMA models and the BSM are quite close, but for one series the signal extraction variances are quite different. Finally, in section 5 we draw some tentative conclusions.

2. ARIMA AND COMPONENTS MODELS

Let Y_t for t=1,...,n be observations on a time series, which will often be the logarithm of some original time series. We write

$$Y_t = \underset{\sim}{X}_t'\underset{\sim}{\beta} + Z_t \tag{2.1}$$

where $\underset{\sim}{X}_t'\underset{\sim}{\beta}$ is a linear regression mean function with $\underset{\sim}{X}_t$ the vector of regression variables at time t and $\underset{\sim}{\beta}$ the vector of regression parameters, and Z_t is the (zero mean) stochastic part of Y_t. The regression variables used here will be to account for trend constants, calendar variation, fixed seasonal effects, and outlier effects (Findley, et. al. 1988). We will be interested in decompositions of Z_t such as

$$Z_t = S_t + N_t = S_t + T_t + I_t \tag{2.2}$$

where S_t is a (stochastic) seasonal component, and N_t a (stochastic) nonseasonal component that can be further decomposed into a trend component T_t and an irregular component I_t. If Y_t is the logarithm of the time series of interest, note (2.1) and (2.2) imply multiplicative decompositions for the original time series.

One approach to analyzing time series components involves modeling Z_t directly, then making assumptions that lead from this model to definitions of and models for the components. The other approach is to directly specify models for the components, which then implies a model for Z_t that can be fitted to data. We shall consider ARIMA models as a basis for both approaches. While other models have certainly received attention in recent years (long memory, ARCH, and nonlinear models come to mind), ARIMA and ARIMA component models seem to have been the most popular, and so focusing attention on these two seems an appropriate starting point.

The ARIMA models we shall use for Z_t can be written in the form (c.f. Box and Jenkins 1976):

$$\phi(B)(1-B)^d(1-B^{12})Z_t = \theta(B)(1-\theta_{12}B^{12})a_t \tag{2.3}$$

where B is the backshift operator ($BZ_t = Z_{t-1}$), $d \geq 0$ (if d=0, $(1-B)^d = 1$), $\phi(B) = 1-\phi_1B-...-\phi_pB^p$ and

$\theta(B) = 1-\theta_1 B-...-\theta_q B^q$ are AR and MA operators of low order (usually p, q $\leq$ 3), and a_t is white noise (iid $N(0,\sigma_a^2)$.) This model is for monthly seasonal data; the modifications for data with other seasonal periods (e.g. quarterly) are obvious, and the $1-B^{12}$ and $1-\theta_{12}B^{12}$ are removed for nonseasonal data. We could have included a seasonal autoregressive operator in (2.3), though we rarely use these. If $\theta_{12} = 1$ we can "cancel" the $1-B^{12}$ factor on both sides of (2.3) and add seasonal mean variables to $\underset{-}{X}_t$ (Abraham and Box 1978, Bell 1987). Identification, estimation, and diagnostic checking of these models proceeds with by now well-established procedures — see Box and Jenkins (1976) for pure ARIMA models, Bell and Hillmer (1983) and Findley et al. (1988) for models with regression terms. Estimation is by maximum likelihood where the likelihood function is defined as the joint density of the differenced data $(1-B)^d(1-B^{12})Y_t$ t=d+13,...,n.

Component models specify simple ARIMA models for the components in (2.2). Harvey and Todd's (1983) basic structural model (BSM) can be written

$$\begin{aligned} Z_t &= S_t + T_t + I_t \\ U(B)\, S_t &= \epsilon_{1t} \qquad \epsilon_{1t} \sim \text{iid } N(0,\sigma_1^2) \\ (1-B)^2 T_t &= (1-\eta B)\epsilon_{2t} \qquad \epsilon_{2t} \sim \text{iid } N(0,\sigma_2^2) \\ I_t &\sim \text{iid } N(0,\sigma_3^2) \end{aligned} \tag{2.4}$$

where $U(B) = 1 + B ... + B^{11}$ sums a series over 12 consecutive months. They actually begin with T_t following a random walk with stochastic drift, where the drift also follows a random walk; this leads to the (0,2,1) model for T_t in (2.4) with the constraint $\eta \geq 0$. While we shall not enforce this constraint, it turns out to be easily satisfied for all our example series here. If the "stochastic" drift has zero innovation variance (i.e. it is actually a constant) then $\eta = 1$ and the model for T_t reduces to $(1-B)T_t = \beta_0 + \epsilon_{1t}$, and we can account for β_0 by adding the time trend variable t to $\underset{-}{X}_t$. If $\sigma_1^2 = 0$ then S_t becomes fixed and can be handled with appropriate variables in $\underset{-}{X}_t$ analogous to what was noted when $\theta_{12} = 1$ in the ARIMA model (2.3).

Gersch and Kitagawa (1983) (see also Kitagawa and Gersch 1984) consider models similar to (2.4), but with T_t following the model

$$(1-B)^\delta T_t = \epsilon_{2t} \qquad \delta = 1,2, \text{ or } 3. \tag{2.5}$$

We whall refer to (2.4) but with T_t following (2.5) as the GK model. Notice that the GK model with $\delta = 2$ becomes the BSM with $\eta = 0$, while the BSM with η=1 is the GK with δ=1 and a trend constant. Akaike (1980) suggested similar models, but with S_t following a model that now seems unattractive.

Gersch and Kitagawa extend their model with the addition of a stationary autoregressive component. This can be written as

$$\begin{aligned} Z_t &= S_t + T_t + I_t + V_t \\ (1-\alpha_1 B - ... - \alpha_p B^p)V_t &= \epsilon_{4t} \qquad \epsilon_{4t} \sim \text{iid } N(0,\sigma_4^2) \end{aligned} \tag{2.6}$$

with S_t and I_t as in (2.4), and T_t as in (2.5). Harvey (1985) also considers such an extension to his models, with the autoregressive parameters constrained so that V_t tends to exhibit cyclical behavior. He also considers an ARMA(2,1) formulation for V_t.

Modeling procedures for these component models are more automatic than for ARIMA models and are discussed in the references cited. Estimation is again by maximum likelihood, with the likelihood evaluated using the Kalman filter. Since the models are nonstationary this presents problems for initialization of the Kalman filter that have been recently addressed by Kohn and Ansley (1986) and Bell and Hillmer (1987a).

These approaches produce a likelihood function that is again the joint density of the differenced data, which is now determined by the components models.

The ARIMA models for the components imply an ARIMA model for the aggregate Z_t, as has been observed by G. C. Tiao (reported in Findley 1983) and Maravall (1985). Taking (2.4) for illustration, applying $(1-B)^2U(B) = (1-B)(1-B^{12})$ to Z_t gives $(1-B)^2\epsilon_{1t} + U(B)(1-\eta B)\epsilon_{2t} + (1-B)(1-B^{12})\epsilon_{3t}$, which follows a moving average model of order 13 whose parameters are determined by σ_1^2, σ_2^2, σ_3^2, and η. While (2.4) is thus equivalent to an ARIMA$(0,1,13)\times(0,1,0)_{12}$ model for Z_t, the high regular MA order and the constraints on the parameters make it unlikely that direct ARIMA modeling of Z_t would yield such a model exactly. Thus, there is potential for difference between the ARIMA and component model approaches, though Maravall (1985) notes that certain parameter values for (2.4) can yield a model close to the popular ARIMA$(0,1,1)\times(0,1,1)_{12}$ "airline model" of Box and Jenkins (1976). For nonseasonal series or series whose seasonality is modeled as fixed through the regression function $\underset{\sim}{X}_t'\underset{\sim}{\beta}$, the ARIMA model implied by (2.4) for $Z_t = T_t + I_t$ depends on $(1-\eta B)\epsilon_{2t} + (1-B)^2I_t$, which follows an MA(2) model whose 3 parameters are determined by σ_2^2, σ_3^2, and η. We could easily get exactly the same model by direct modeling of Z_t as ARIMA (0,2,2). Similar results obtain for other nonseasonal components models. While the potential for difference between nonseasonal ARIMA and components models is difficult to judge, the potential for ARIMA and components models to be effectively the same seems greater in the nonseasonal than in the seasonal case.

This discussion raises questions about how much ARIMA and components models will differ in practice, and which will fit better when they do differ? We will make a preliminary investigation into this by comparing the fit of ARIMA and components models on a set of time series. As the models we wish to compare are generally nonnested (one is not obtained by simple constraints on the parameters of the other) traditional hypothesis tests or confidence intervals would be difficult to apply. We shall use the AIC criterion of Akaike (1973), which is defined as

$$\text{AIC} = -2\hat{L} + 2m$$

where $\hat{L}$ is the maximized log–likelihood and m is the number of parameters estimated. The model with the smaller AIC is to be preferred. To compare two models, 1 and 2 say, we present the difference in their AIC's, DAIC = $\text{AIC}_1 - \text{AIC}_2$. A positive value of DAIC favors model 2, a negative value model 1. Judging when there is a "significant" difference between models as measured by DAIC is not necessarily straightforward (see Findley 1988), but users of AIC often view differences of 1 or 2 as significant. We shall use 2 as a rough significance boundary. As a crude justification, notice that if we add a parameter to a model $\hat{L}$ cannot decrease, so if the parameter yields no improvement in fit, $\hat{L}$ remains the same and AIC increases by 2.

We shall use AIC to compare the fit of ARIMA and components models on a set of Census Bureau seasonal time series analyzed by Burman and Otto (1988) using ARIMA models. (Many were analyzed previously in Hillmer, Bell, and Tiao (1983), though with fewer years of data available. We also include one series, labelled ENM20, from the U.S. Bureau of Labor Statistics, analyzed in Bell and Hillmer 1984.) These series have the advantage of having readily available models with careful treatment of regression terms for calendar variation, fixed seasonal effects (occasionally), and outliers. We exclude a few series Burman and Otto (1988) analyzed that are not published, as well as the foreign trade series they analyzed since these have undergone significant revisions in recent years to correct some major data problems. This leaves 45 series for analysis, listed in Bell and Pugh (1990). The series are broadly representative of the series seasonally adjusted by the Census Bureau, but are not a random sample, so the analysis here might be best viewed as a pilot study.

For a given series we shall use the same regression terms with both ARIMA and components models, and also will restrict comparisons to models with the same order of differencing. Comparing models with different orders of differencing poses some problems since the likelihood functions for the two models are then based on different (differenced) data. This restriction means that we will compare ARIMA models (2.3) with d=1 to the BSM as in (2.4). ARIMA models with d=0 will be compared to a model as in (2.4), but with T_t following (2.5) with δ=1. Models with a fixed seasonal and d=1 in the ARIMA structure will be compared to a components model with a fixed seasonal (no stochastic S_t), and with T_t again following (2.5) with δ=1. The latter two cases correspond to particular cases of both the BSM and GK models. When the ARIMA model has d=1 and a stochastic seasonal, we shall not make comparisons with the GK model that would use (2.5) with δ=2. Since this is a special case of (2.4) with η=0, at best this GK model would avoid one extraneous parameter and have an AIC 2 less than that of (2.4). At worst, it can have a substantially higher AIC than (2.4) if the maximum likelihood estimator $\hat{\eta}$ is not near 0 (though if $\hat{\eta} \approx 1$ we can think of (2.4) as overdifferencing the GK model with δ=1.)

The ARIMA models used and their AICs, the fitted BSMs and their AICs, and the AIC differences are given in Bell and Pugh (1990). Table 1 below provides a summary. The results are obvious: AIC exhibits a strong preference for ARIMA models overall, with large AIC differences (> 8) for about one half of the series. DAIC's for the two series for which the BSM was preferred were only –2.1 and –2.7.

Table 1: BSM versus ARIMA

# series in	Order of Differencing		
DAIC range	(1,1)	(0,1)	(1,0)
< – 2	2	0	0
–2 to 2	6	1	0
2 to 8	9	2	3
8 to 20	10	3	2
20 to 40	5	0	1
> 40	4	0	0
	36	6	6

(Three series appear twice in the table since they were refit with fixed seasonals after getting $\hat{\theta}_{12} \approx 1$.)

In looking for possible explanations for the poor fit of the BSM we examined DAICs and corresponding $\hat{\theta}_{12}$'s, $\hat{\eta}$'s, etc., but found no obvious patterns. Selection bias was considered as a possible explanation, even though the ARIMA models were selected with the usual identification approach based on autocorrelations and partial autocorrelations, and not by searching a set of models for the model with minimum AIC. To check for selection bias, the BSM AICs were compared with those for the ARIMA$(0,1,1)\times(0,1,1)_{12}$ "airline model", which seems a reasonable choice if one were to use a single ARIMA model. Although the BSM fit much better than the airline model for two series (DAICs of –11.7 and –25.6), aside from this the results changed little from those in Table 1. This is perhaps not surprising since 15 of the selected ARIMA models were airline models, and others were not very different from the airline model. The airline model performed much better in comparison to the selected ARIMA models than the BSM, though four series favored the selected ARIMA model over the airline model by an AIC greater than 20, suggesting that use of any single model for all series will occasionally lead to poor fits.

This report would not be complete without some comments on our experience fitting components models. The results presented here were obtained using a computer program for fitting time series models with ARIMA components and regression terms recently developed by ourselves, other members of the Time Series Staff of the Statistical Research Division at Census, and Steven Hillmer of the University of Kansas. We found the components models much more difficult to fit than regular ARIMA models. For example, getting good starting values for nonlinear iteration over the component model parameters seems important, whereas we find getting good starting values for ARIMA model parameters not at all important. We have not presented results for models with a fourth component as in (2.6) because we were unable to successfully fit such models. Adding a fourth component casued the nonlinear search to go outside the stationarity region for V_t, causing the program to crash on every series. While there are means of programming around this problem, and while inclusion of a fourth component might improve the fits, we found these difficulties discouraging. Though we did not make a formal study of the numerical problems we experienced with components models, they seemed due to the likelihood being rather flat in certain directions in the parameter space. Given this, we find the oft–claimed advantages of "simplicity" and "interpretability" for components models difficult to accept.

The computational difficulties we experienced suggest a final possible explanation for our results – that there is something wrong with our software and it is not actually maximizing the likelihood. While we have checked our program thoroughly, and do not believe this to be the case, we cannot rule this out with certainty. We will gladly provide our data to anyone interested in checking our results. We would be even more interested in seeing a study done with other series to see if similar results are obtained.

3. SEASONAL ADJUSTMENT

While section 2 suggests that ARIMA models may fit a time series substantially better than components models, there is still the question of what difference choice of a model makes in practice? Here we consider the effect of model choice on seasonal adjustment. For a given components model, seasonal adjustment can be done by applying a Kalman smoother to the series (see, e.g., Gersch and Kitagawa 1983). With ARIMA models one must first make sufficient assumptions leading from simple ARIMA models for observed series to unique component models. This is addressed by Burman (1980) and Hillmer and Tiao (1982), who consider a range of possible decompositions and suggest a choice leading to a unique decomposition into component models. (The two approaches differ some for certain models that do not seem to occur often.) The underlying assumptions are set out and discussed further by Bell and Hillmer (1984). As will be seen shortly, we can also consider a range of decompositions for any given components model.

For Y_t following (2.1) and (2.3), Burman (1980) and Hillmer and Tiao (1982) achieve a decomposition of form (2.2) by making a partial fractions decomposition of the covariance generating function (CGF), $\gamma_Z(B)$, of Z_t, yielding CGF's $\gamma_S(B)$, $\gamma_T(B)$, and $\gamma_I(B)$, and corresponding ARIMA models for the components. This yields a range of admissible decompositions corresponding to $\gamma_Z(B) = [\gamma_S(B) - \gamma_1] + [\gamma_T(B) - \gamma_2] + [\gamma_I(B) + \gamma_1 + \gamma_2]$, for any γ_1 and γ_2 such that each bracketed term is ≥ 0 for all $B = e^{i\lambda}$. The range reflects inherent uncertainty about the decomposition; specifying γ_1 and γ_2 yields a particular decomposition that can be used for seasonal adjustment. Burman (1980) and Hillmer and Tiao (1982) suggest picking the maximum possible γ_1 and γ_2 ($\bar{\gamma}_1 = \min_\lambda \gamma_S(e^{i\lambda})$ and $\bar{\gamma}_2 = \min_\lambda \gamma_T(e^{i\lambda})$), leading to what is called the canonical decomposition, which has several attractive properties. Focusing in particular on the seasonal–nonseasonal decomposition now, the components corresponding to any admissible γ_1 can be written as $S_t = \bar{S}_t + \nu_t$ and $N_t = \bar{N}_t - \nu_t$, where $\bar{S}_t$ and $\bar{N}_t$ are the canonical seasonal and nonseasonal, and ν_t is white noise with variance $\bar{\gamma}_1 - \gamma_1$. Thus, the canonical decomposition can be viewed as removing as much white noise as possible from the seasonal component and putting it in the nonseasonal through the irregular. Since there is no apparent reason to include additional white noise in the seasonal, this is a good argument for using the canonical decomposition. (Watson (1987) gives an approach that avoids assuming a particular decomposition.)

(As an aside, we note that it is also necessary to decompose the deterministic regression effects, $\underline{X}_t'\underline{\beta}$, into seasonal and nonseasonal parts. This is discussed in Bell (1984), but since there is no reason to do this differently for ARIMA and components models we need not go into it here.)

Bell and Hillmer (1984) criticize component modelers for simply taking the component models for adjustment as those obtained in modeling the observed series, and thus ignoring the uncertainty inherent in the basic decomposition into components. We can address this decomposition uncertainty for component models by defining a "canonical decomposition" in an analogous way to that defined for ARIMA models — subtracting as much white noise as possible from S_t and adding it to N_t through I_t. In Bell and Pugh (1990, Appendix A.1) we show that the resulting canonical components model decomposition, $Z_t = \bar{S}_t + \bar{N}_t = \bar{S}_t + [T_t + \bar{I}_t]$, has a canonical irregular $\bar{I}_t$ with variance $\bar{\sigma}_3^2 = \sigma_3^2 + \sigma_1^2/144$, and a canonical seasonal $\bar{S}_t$ which follows the model

$$U(B)\bar{S}_t = \psi(B)\bar{\epsilon}_{1t} \qquad \bar{\epsilon}_{1t} \sim \text{iid } N(0,\bar{\sigma}_1^2) \tag{3.1}$$

where $\psi(B)$, of order 11, is given in Table 2., and $\bar{\sigma}_1^2 = .8081\ \sigma_1^2$. (Bell and Pugh (1990) also discuss a

Table 2.: Coefficients ψ_k for $\psi(B) = 1 - \psi_1 B - \cdots - \psi_{11}B^{11}$

k	ψ_k	k	ψ_k	k	ψ_k
1	.205555	5	.100648	9	.031188
2	.175919	6	.080059	10	.018953
3	.148557	7	.061661	11	.008593
4	.123471	8	.045395		

canonical trend for components models.) This is in fact the same form as the canonical seasonal model of Burman (1980) and Hillmer and Tiao (1982), though their seasonal model will generally have a different $\psi(B)$ and $\bar{\sigma}_1^2$ (that depend on the ARIMA model). As with ARIMA models, using any other admissible decomposition (corresponding to any valid decomposition of the covariance generating function), including that defined by the original fitted components model, can be viewed as adding white noise to the canonical seasonal $\bar{S}_t$. Notice that, given a components model, the model for $\bar{S}_t$ in (3.1) is trivial to obtain. Also, signal extraction for canonical seasonal adjustment may be performed in the usual way with a Kalman smoother using the model (3.1) for $\bar{S}_t$ and increasing the irregular variance to $\bar{\sigma}_3^2$.

Notice that the amount of variance removed from the components model seasonal, $\sigma_1^2/144$, will be small unless σ_1^2 is large relative to σ_2^2 and σ_3^2. However, the opposite is true for the series considered here: $\sigma_1^2/(\sigma_2^2 + \sigma_3^2)$ exceeds .07 for only two of the 45 series. This has two implications: (1) the estimated component model typically implies a very nearly fixed seasonal, and (2) the original component model decomposition will often be very close to the canonical component model decomposition. In fact, for the examples we have tried, seasonal adjustments from the original and canonical component model decompositions have been virtually identical. Since this aspect of decomposition choice appears to make little difference we shall not consider it further here. This is not to say choosing some other decomposition than the canonical cannot have important effects, though we shall not consider that here either.

To examine potential differences in seasonal adjustments arising from model choice we examine seasonal adjustments for two series: IHAPVS (value of U.S. household appliances shipped from 1/62–12/81), and ENM20 (thousands of employed males 20 and older in nonagricultural industries from 1/65 – 8/79), a series analyzed by Bell and Hillmer (1984). IHAPVS was one of the series which the BSM fit best (DAIC = –.7), while the BSM fit for ENM20 was rather poor (DAIC = 13.7), though far from the worst. ENM20 was the one series for which logarithms were not taken so an additive decomposition is used here.

Figure 1.a. shows the estimated ARIMA and BSM seasonal components for IHAPVS. Close inspection is required to detect any difference. As this is also true of the seasonal adjustments we do not present these. Figure 1.b. shows the signal extraction standard deviations for IHAPVS expressed as coefficients of variation. Here substantial differences appear with the ARIMA CV's being 20 percent or more higher near the end of the series. (Note the results for the ARIMA model are not necessarily bad.) However, the CV's might all be considered small: none exceed about 1.6 percent.

Figure 2.a shows the ARIMA and BSM seasonals for ENM20. Here we can see a difference: the ARIMA seasonal evolves steadily over time while the BSM seasonal remains relatively fixed. (For ENM20 the BSM has $\hat{\sigma}_1^2 = 27$ and $\hat{\sigma}_2^2 = 16{,}500$.) Figure 2.b portrays seasonal adjustment results for the last 5 years of the data. While differences can be seen they may not be important since the month–to–month changes themselves are not large, seldom exceeding .5 percent. Figure 2.c. shows even larger differences for signal extraction standard deviations than we saw for IHAPVS. The BSM standard deviations rise very little at the end of the series because an essentially fixed seasonal is being estimated. Still, the most noteworthy aspect of Figure 2.c. may be how small the standard deviations are relative to series values of 40,000 to 50,000.

We conjecture that $\text{Var}(S_t - \hat{S}_t) \longrightarrow 0$ as $\theta_{12} \longrightarrow 1$ in the ARIMA model and as $\sigma_1^2 \longrightarrow 0$ in the BSM, which probably explains the small signal extraction standard deviations observed in the two examples. However, if we decide $\theta_{12} = 1$ or $\sigma_1^2 = 0$ and use a model with fixed seasonal regression effects instead, the signal extraction variances will not be 0 since we will have error in estimating the seasonal regression parameters. A curious aspect of these results is the apparent discontinuity between results for $\theta_1 < 1$ (or $\sigma_1^2 > 0$) and $\theta_{12} = 1$ (or $\sigma_1^2 = 0$).

4. REPEATED SURVEY ESTIMATION

Scott and Smith (1974) and Scott, Smith and Jones (1977) suggested using time series signal extraction techniques for estimation in periodic surveys. If s_t denotes the true population quantity (the signal) and e_t the sampling error at time t, then we use signal extraction to estimate s_t in

$$Y_t = s_t + e_t\,, \tag{4.1}$$

If Y_t is the logarithm of the original series, then $\exp(s_t)$ and $\exp(e_t)$ are the true population quantity and multiplicative sampling error in the original series. Any of the models discussed in section 2 can be used for s_t; Binder and Dick (1989) and Bell and Hillmer (1989) use ARIMA models, while Pfefferman (1989) uses a BSM. Generally, any regression terms in the model are also part of s_t.

Model building for the survey estimation problem is discussed in the references cited above. A primary distinction between this application and what we have considered before, is that the model for e_t is generally estimated, in some fashion, using survey microdata. The sampling error model is then held fixed when estimating the parameters of the s_t model using the time series data on Y_t. Questions arise about the sensitivity of the survey estimation results to any of the aspects of the modeling. Here we shall examine the sensitivity of results to the choice between an ARIMA model and a BSM for s_t.

We consider two time series. For the first, U.S. teenage unemployment (in 1000's) from 1/72 to 12/83, Bell and Hillmer (1987b) develop the following model for $Y_t = s_t + e_t$:

$$(1-B)(1-B^{12})s_t = (1-.27B)(1-.68B^{12})a_t \qquad \sigma_a^2 = 4294$$

$$e_t = h_t\tilde{e}_t \qquad (1-.6B)\tilde{e}_t = (1-.3B)c_t \qquad \sigma_c^2 = .8767 \qquad h_t^2 = -.0000153\,Y_t^2 + 1.971\,Y_t$$

The model for s_t has been reestimated, yielding slightly different parameter values than those reported in Bell and Hillmer (1987b). With $\sigma_c^2 = .8767$, $\text{Var}(\tilde{e}_t) = 1$, so h_t is the (estimated) sampling error standard deviation, which is time–varying. The modeling of the second series, U.S. 5 or more unit housing starts, is very similar to that for U.S. single family housing starts, also considered in Bell and Hillmer (1987b). The sampling errors for this series appear approximately uncorrelated over time with relative variance .00729, which is also the approximate variance of the logged multiplicative sampling errors. The estimated ARIMA model for the signal in the logged time series is

$$(1-B)(1-B^{12})s_t = (1-.47B)(1-.89B^{12})a_t \qquad \sigma_a^2 = .0215.$$

We used the above models in signal extraction estimation of s_t, and then did the same with a BSM fitted for s_t with the same e_t models given above. The BSM model fitted relatively well for both these series, with DAIC = AIC(BSM) – AIC(ARIMA) = – 3.1 for teenage unemployment and DAIC = 1.8 for housing starts. (The appropriateness of these AICs is in some question since the e_t models are not fitted with the time series data.) Figure 3.a. shows the signal extraction point estimates for teenage unemployment using both models; $(1-B^{12})\hat{s}_t$ is shown to avoid the obscuring effects of seasonality. The BSM estimates less variance in the signal than the ARIMA model, and thus yields slightly smoother estimates. Figure 3.b. shows substantial differences in the signal extraction variances for the two models. The two signal extraction estimates for the housing starts series were virtually identical, and so are not shown. Figure 4 shows the signal extraction coefficients of variation (standard deviations for the logged series) for the last half of the housing starts series — those for the first half would be a mirror image. While there are some interesting differences in pattern, the magnitude of the differences is small.

5. CONCLUSIONS

Even the conclusions drawn in section 2 must be somewhat tentative; it would be interesting to see similar studies with other sets of time series. Because of the limited examples considered in sections 3 and 4, the conclusions there can only be suggestive. To summarize:

1. Data can frequently discriminate between ARIMA and components models. For the 45 series analyzed, AIC showed a strong general preference for ARIMA models over the BSM. To the extent that model fit is important, merely assuming the BSM provides an adequate fit could be dangerous.

2. We found fitting components models more difficult than fitting ARIMA models. While we would have liked to see if the addition of a stationary AR component or other cycle term could improve the component model fits, we were unable to fit such models due to numerical problems.

3. Signal extraction point estimates for seasonal adjustment and survey estimation using ARIMA models and using the BSM differed little for the examples considered. Signal extraction variances showed much larger differences, though for the seasonal adjustment examples the variances using both models might be regarded as quite small. This last point is worth more investigation, to see if model–based seasonal adjustment variances with canonical, or approximately canonical, decompositions are typically very small.

REFERENCES

Abraham, B. and Box, G.E.P. (1978), "Deterministic and Forecast–Adaptive Time–Dependent Models," Applied Statistics, 27, 120–130.

Akaike, H. (1973) "Information Theory and an Extension of the Likelihood Principle," in the 2nd International Symposium on Information Theory, eds. B. N. Petrov and F. Czaki, Budapest: Akademia Kiado, 267–287.

_______ (1980), "Seasonal Adjustment by a Bayesian Modeling," Journal of Time Series Analysis, 1, 1–13.

Bell, W. R. (1984) "Seasonal Decomposition of Deterministic Effects," Research Report Number 84/01, Statistical Research Division, Bureau of the Census.

Bell, W. R. (1987) "A Note on Overdifferencing and the Equivalence of Seasonal Time Series Models With Monthly Means and Models With $(0,1,1)_{12}$ Seasonal Parts When $\Theta = 1$," Journal of Business and Economic Statistics, 5, 383–387.

Bell, W. R. and Hillmer, S. C. (1983), "Modeling Time Series with Calendar Variation," Journal of the American Statistical Association, 78, 526–534.

_______ (1984), "Issues Involved with the Seasonal Adjustment of Economic Time Series," (with discussion), Journal of Business and Economic Statistics, 2, 291–320.

_______ (1987a), "Initializing the Kalman Filter for Nonstationary Time Series Models," Research Report Number 87/33, Statistical Research Division, Bureau of the Census.

_______ (1987b), "Time Series Methods for Survey Estimation," Research Report Number 87/20, Statistical Research Division, Bureau of the Census.

_______ (1989), "Modeling Time Series Subject to Sampling Error," Research Report Number 89/01, Statistical Research Division, Bureau of the Census.

Bell, W. R. and Pugh, M. G. (1990) "Alternative Approaches to the Analysis of Time Series Components," Research Report Number 90/01, Statistical Research Division, Bureau of the Census.

Binder, D. A. and Dick, J. P. (1989) "Modelling and Estimation for Repeated Surveys," Survey Methodology, 14, to appear.

Box, G.E.P. and Jenkins, G. M. (1976), Time Series Analysis: Forecasting and Control, San Francisco: Holden Day.

Burman, J. P. (1980), "Seasonal Adjustment by Signal Extraction," Journal of the Royal Statistical Society Series A, 143, 321–337.

Burman, J. P. and Otto, M. (1988), "Outliers in Time Series," Research Report Number 88/14, Statistical Research Division, Bureau of the Census.

Buys Ballot, C. H. D. (1847) Les Changements Periodiques de Temperature, Utrecht: Kemink et Fils.

Carlin, J. B., and Dempster, A. P. (1989) "Sensitivity Analysis of Seasonal Adjustments: Empirical Case Studies," Journal of the American Statistical Association, 84, 6–20.

Findley, D. F. (1983), "Comments on 'Comparative Study of the X–11 and BAYSEA Procedures of Seasonal Adjustment' by H. Akaike and M. Ishiguro," in Applied Time Series Analysis of Economic Data, ed. Arnold Zellner, Washington, D.C.: U. S. Department of Commerce, Bureau of the Census.

_______ (1988), "Comparing Not Necessarily Nested Models With the Minimum AIC and the Maximum Kullback–Leibler Entropy Criteria: New Properties and Connections," Research Report Number 88/21, Statistical Research Division, Bureau of the Census.

Findley, D. F., Monsell, B. M., Otto, M. C., Bell, W. R., and Pugh, M. G. (1988) "Toward X–12 ARIMA," Proceedings of the Fourth Annual Research Conference, U. S. Department of Commerce, Bureau of the Census.

Gersch, W. and Kitagawa, G. (1983), "The Prediction of Time Series With Trends and Seasonalities," Journal of Business and Economic Statistics, 1, 253–264.

Harvey, A. C. (1985), "Trends and Cycles in Macroeconomic Time Series," Journal of Business and Economic Statistics, 3, 216–227.

Harvey, A. C. and Todd, P. H. J. (1983), "Forecasting Economic Time Series With Structural and Box–Jenkins Models: A Case Study," (with discussion), Journal of Business and Economic Statistics, 1, 299–315.

Hillmer, S. C., Bell, W. R., and Tiao, G. C. (1983a), "Modeling Considerations in the Seasonal Adjustment of Economic Time Series," in Applied Time Series Analysis of Economic Data, ed. Arnold Zellner, U.S. Department of Commerce, Bureau of the Census, 74–100.

Hillmer, S. C., and Tiao, G. C. (1982), "An ARIMA–Model–Based Approach to Seasonal Adjustment," Journal of the American Statistical Association, 77, 63–70.

Hotta, L. K. (1989), "Identification of Unobserved Components Models," Journal of Time Series Analysis, 10, 259–270.

Kitagawa, G. and Gersch, W. (1984), "A Smoothness Priors–State Space Modeling of Time Series With Trend and Seasonality," Journal of the American Statistical Association, 79, 378–389.

Kohn, R. and Ansley, C. F. (1986), "Estimation, Prediction, and Interpolation for ARIMA Models With Missing Data," Journal of the American Statistical Association, 81, 751–761.

Maravall, A. (1985), "On Structural Time Series Models and the Characterization of Components," Journal of Business and Economic Statistics, 3, 350–355.

Nerlove, M., Grether, D. M., and Caravallo, J. L. (1979), Analysis of Economic Time Series: A Synthesis, New York: Academic Press.

Prothero, D. L. and Wallis, K. F. (1976) "Modeling Macroeconomic Time Series," Journal of the Royal Statistical Society Series A, 139, 468–500.

Pfeffermann, D. (1989) "Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys," paper presented at the annual meeting of the American Statistical Association, Washington, D. C.

Scott, A. J. and Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods," Journal of the American Statistical Association, 69, 674–678.

Scott, A. J., Smith, T.M.F., and Jones, R. G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," International Statistical Review, 45, 13–28.

Watson, M. W. (1987) "Uncertainty in Model–Based Seasonal Adjustment Procedures and Construction of Minimax Filters," Journal of the American Statistical Association, 82, 395–408.

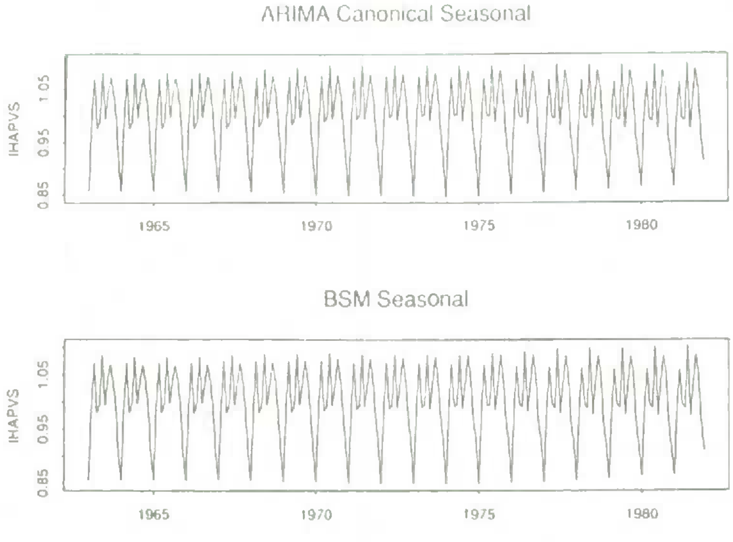


Figure 1.a

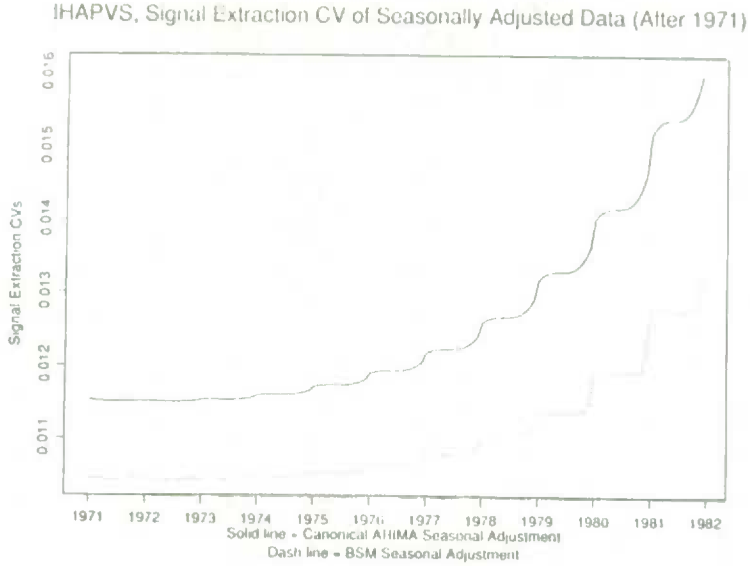
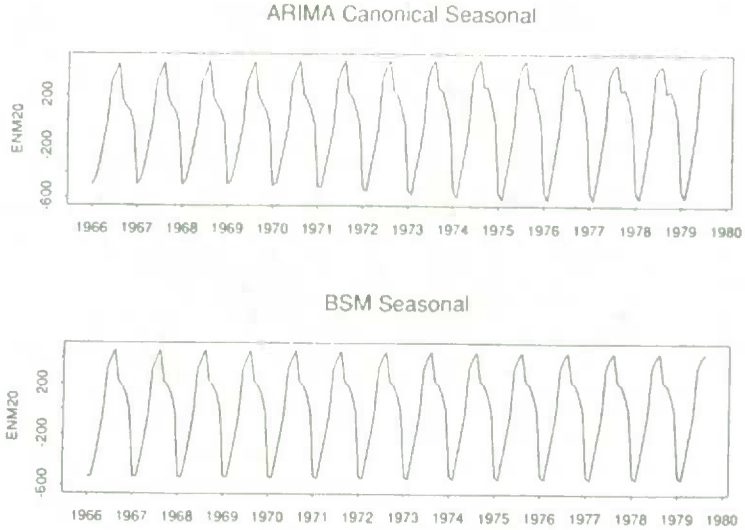

IHAPVS, Signal Extraction CV of Seasonally Adjusted Data (After 1971)

Figure 1.b


ARIMA Canonical Seasonal

BSM Seasonal

Figure 2.a

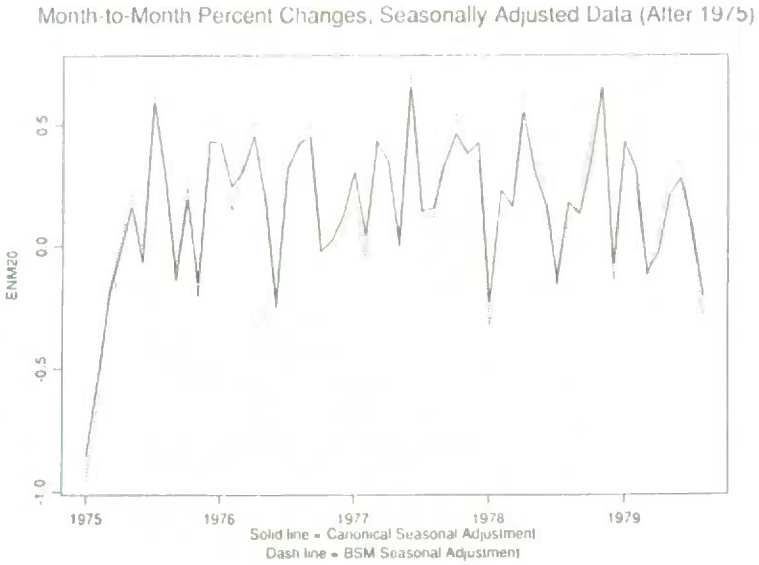
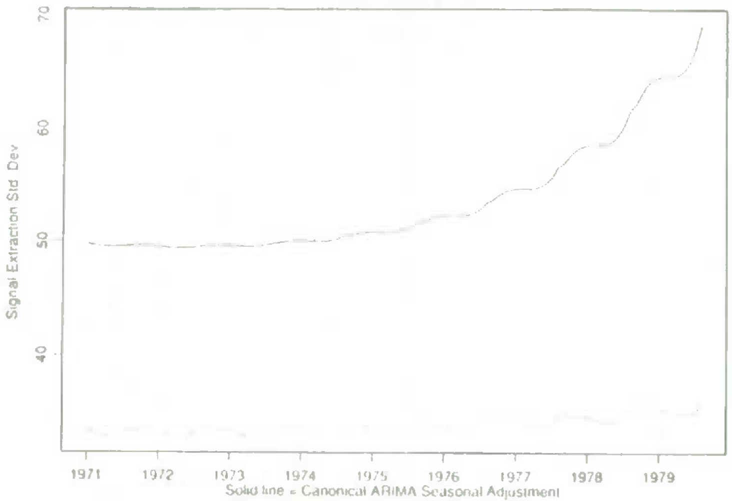

Month-to-Month Percent Changes, Seasonally Adjusted Data (After 1975)

Figure 2.b

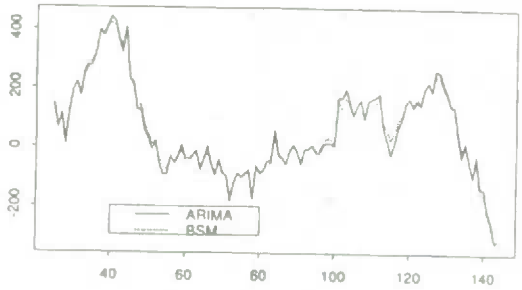
ENM20, Signal Extraction Std. Dev. of Seasonally Adjusted Data (After 1971)



Solid line = Canonical ARIMA Seasonal Adjustment
Dash line = BSM Seasonal Adjustment

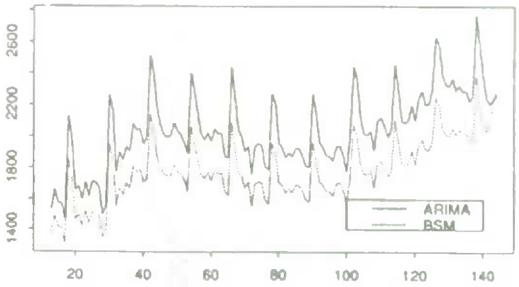
Figure 2.c

Teenage Unemployment



Signal Extraction Estimates

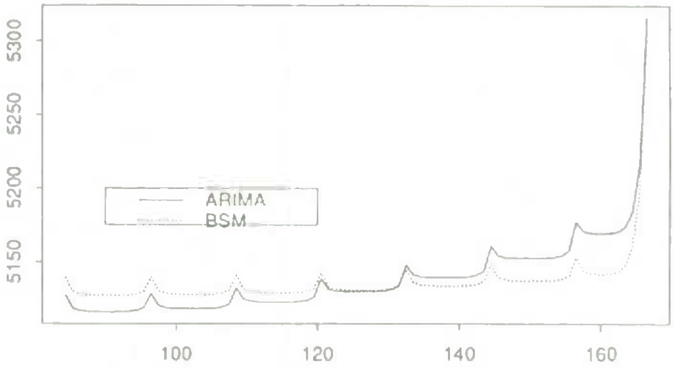
Figure 3.a



Signal Extraction Variances

Figure 3.b

Total US 5+ Housing Starts



Signal Extraction Variances

Figure 4.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

REG-ARIMA BASED PREPROCESSING FOR SEASONAL ADJUSTMENT

D.F. Findley and B.C. Monsell[1]

ABSTRACT

The time series staff of the Census Bureau's Statistical Research Division has developed software modules which can be adapted to existing seasonal adjustment programs to provide pre- and postprocessing for enhanced adjustment and quality control capabilities. The preprocessing module is a program for modeling and doing computationally efficient "exact" maximum likelihood estimation of seasonal ARIMA models with a regression mean function. Many regressors are built into the software, to permit the user to detect and model a variety of common outlier and calendar effects which occur in economic data and which existing seasonal adjustment programs either cannot treat or frequently do not handle well. The program also allows the user to include their own regressor variables. This note presents some examples illustrating the use of the preprocessing module.

KEY WORDS: REG-ARIMA Model; AIC.

1. INTRODUCTION

For many economic time series, establishing an appropriate seasonal adjustment procedure requires several cycles of preadjustment and postadjustment processing. The preprocessing involves forecast extensions and data adjustments which are performed, perhaps tentatively, before the actual seasonal adjustment moving averages are applied to the series. Postprocessing refers to the calculation of a variety of diagnostics to evaluate the effects on the seasonally adjusted series of the preprocessing and adjustment options which were chosen. The main goal of postprocessing is to determine if a satisfactory adjustment has been achieved. We have developed a new set of techniques for postprocessing, called sliding spans analysis, which is described in Findley, Monsell, Shulman and Pugh (1990).

This note concerns preprocessing. We present four examples demonstrating the valuable role of what we shall call REG-ARIMA (regression + ARIMA) models for determining or comparing preadjustments. Capabilities for identifying and estimating both typical and customized REG-ARIMA models are included in the preprocessing module of a seasonal adjustment program, provisionally called X-12-ARIMA, which is nearing completion at the U.S. Census Bureau, see Findley, Monsell, Otto, Bell and Pugh (1988). This program also calculates the sliding spans diagnostics.

2. REG-ARIMA MODELS

Many economic time series show occasional large erratic movements over a short time interval which are preceded and followed by longer periods of reasonably stable fluctuations. Such disruptions can be caused by external events such as strikes, extreme weather conditions, international hostilities and changes in government policies, or they can result from internal factors such as changes in the economic classification scheme or the sample used to define or obtain the series. Such disruptions, especially those which result in a long-lasting change in the level of the series, compromise the reliability of seasonal adjustments obtained from X-11-ARIMA and related procedures, and they also make it difficult to identify ARIMA models for forecasting such series.

Frequently it is possible to model these disruptions adequately by means of REG-ARIMA models, which we will now describe. Let x_t denote the series to be modeled (often the logarithm of the observed series y_t), let B denote the backshift operator, $Bx_t = x_{t-1}$, and let z_t denote a vector of known regression variables whose coefficient vector β can contain both known and unknown coefficients. The unknown coefficients will be calculated as a subvector of the maximum Gaussian likelihood estimates of the unknown parameters of a REG ARIMA model, meaning a time series model of the form

$$\phi(B)(x_t-\beta z_t) = \theta(B)a_t, \tag{2.1}$$

where $\phi(B)$ and $\theta(B)$ are polynomials having no roots with magnitude less than one, and a_t is a white noise process uncorrelated with preceding values of x_t. Our method for estimating such models is described in Findley et al. (1988). If $\hat{L}_N$ denotes the maximized value of the log-likelihood function from N observations

[1] Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

$x_1, \ldots, x_N$ and if total number of coefficients estimated in $\phi(B)$, $\theta(B)$ and β is p, then Akaike's AIC comparison statistic for the fitted model is defined as

$$AIC_N = -2\hat{L}_N + 2p \cdot$$

When two or more estimated models are being compared, the model with smaller AIC_N is usually preferred, see Brockwell and Davis (1988) and Findley (1988), for example. (The theory only supports such comparisons via AIC when the $\phi(B)$ polynomials in all the models have the same number of roots with magnitude 1.)

We list below seven typical sets of regression variables which might be included in z_t and which are available in the preprocessing module of X-12-ARIMA.

1. Additive Outlier at t_0

$$AO_t^{(t_0)} = \begin{matrix} 1, & t=t_0 \\ 0, & t \neq t_0 \end{matrix}$$

2. Level Shift at t_0

$$LS_t^{(t_0)} = \begin{matrix} 1, & t \geq t_0 \\ 0, & t < t_0. \end{matrix}$$

3. Ramp Between t_0 and t_1

$$R_t^{(t_0,\ t_1)} = \begin{matrix} 1 & , t \geq t_1 \\ (t-t_0)/(t_1-t_0), & t_0 < t < t_1 \\ 0 & , t \leq t_0 \end{matrix}$$

4. Preadjustment Divisor for Observed Series y_t

Assuming $x_t = \log(y_t)$ and D_t is a positive number to be divided into y_t (for example, a deflator or a user-defined estimate of the effect of a special short term campaign to promote sales), we define

$$d_t = \log D_t$$

and set the corresponding regression coefficient in β equal to 1, to obtain

$$x_t - d_t = \log(y_t/D_t).$$

5. Monthly Trading Day Variables

If $MW_t^{(j)}$ denotes the number of week days of type j in month t, with $j=1,\ldots,7$ designating Monday,..., Sunday respectively, then we define

$$MTD_t^{(j)} = MW_t^{(j)} - MW_t^{(7)} \quad , j=1,\ldots,6.$$

6. Leap Year February Variable

$$LYF_t = \begin{matrix} -.25 & \text{in a non-leap year February} \\ .75 & \text{in a leap year February} \\ 0 & \text{otherwise,} \end{matrix}$$

see Bell and Hillmer (1983) and Bell (1984).

7. Fixed Seasonal Variables

Let m be the number of periods in the year in which an observation is obtained. (Thus, m=12 for monthly data and m=4 for quarterly data.) Let $I_t^{(j)}$ be the indicator variable for the j-th period, j=1, ..., m. (For example, if m=4, then $I_t^{(j)} = 1$ if y_t is the datum for the j-th quarter of some year, and $I_t^{(j)} = 0$ otherwise). Then we define

$$FS_t^{(j)} = I_t^{(m)} - I_t^{(4)}, \ j=1, \ldots, m-1.$$

The program also includes regression variables for the effect of Easter on retail sales and for the effects of several other U.S. holidays. Users can input their own regression variables for other special effects or other lunar calendar holidays that move between several solar calendar months and have an economic impact, such as Ramadan or the Chinese New Year.

One special set of regressors we considered recently were used to estimate quarterly trading day effects.

<u>Quartely Trading Day and Leap Year First Quater Variables</u>

Let $QW_t^{(j)}$ denote the number of weekdays of type j (as in 5.) in quarter t. We define

$$QTD_t^{(j)} = QW_t^{(j)} - QW_t^{(7)} \quad , 1 \leq j \leq 6 \ .$$

The leap-year first-quarter regressor $LYQ1_t$ is defined by replacing Februaries in the definition of LYF_t in 6. above with first quarters.

3. EXAMPLES

We now present some REG-ARIMA model-based analyses which utilize the variables defined in the preceding section.

Akaike's minimum AIC procedure described above will be used when two competing models must be compared. When model 1 is a restricted form of model 2 with fewer parameters to be estimated, this procedure has a conventional interpretation: a test of hypotheses could be done under the null hypothesis that model 1 is correct by assuming the chi-square asymptotic distribution of the log-likehood ratio,

$$H_0: 2(\hat{L}_N^{(2)} - \hat{L}_N^{(1)}) \sim \chi^2(d),$$

which leads to $AIC_N^{(1)} - AIC_N^{(2)} \sim \chi^{(2)}(d) - 2d$. As a consequence, the condition

$$AIC_N^{(1)} - AIC_N^{(2)} \geq 1$$

would usually be interpreted as a statistically significant difference in AIC values, favoring model 2 (rejection of H_0).

3.1 <u>Change of Definition of Series.</u>

As part of a U.S. government program to reduce the burden on firms of responding to government surveys, a law was changed to require fewer companies to respond to the survey conducted for the Quarterly Financial Report, beginning in the first quarter of 1982. As a result, the levels of some of the series dropped sharply in a way that the trend estimation procedures in the X-11-ARIMA program could not adequately follow, see Figures 1 and 2 below. An additional concern is that the post-1981 segment of the series might have a different seasonal pattern from the pre-1982 segment because of the changed sample. To investigate this possiblitity, two competing REG-ARIMA models were fit to these series. These contained in their regression variables a level-shift at 1982/1 and either a single set of fixed seasonal variables for the full series (model 1) or two sets of such variables (model 2), one set for the segment 1974/1 - 1981/4 and the other for the remainder of the series. This means that for model 2, the coefficients for the seasonal effect before and after the level shift can be different. Thus, if the AR and MA lags in the fitted models are the same, then model 1 is a restricted form of model 2 obtained by requiring the two sets of fixed seasonal variables in model 2 to have identical coefficients. The use of fixed seasonals is a device to permit us to use model comparisons to decide if the seasonal pattern of more recent data must be estimated using only post 1981 data. Table 1 gives the AIC values for REG-ARIMA models with these two types of regressor variables fit to the series of Net Income from Retail Sales (NRS, see Fig. 1) and Net Wholesale Trade Income After Taxes (NWTAT, see Fig. 2). For both comparisons, the difference d in the number of estimated variables is 3.

Table 1. AIC Values Testing for a Changed Seasonal Pattern.

	Same Fixed Seasonals (model 1)	Different Fixed Seasonals (model 2)
NRS	1028.8	993.1
NWTAT	757.0	760.0

Thus, as Fig. 1 suggests, there is a significant change in the seasonal component of NRS in 1982, but not in the seasonal component of NWTAT.

3.2 Testing for the Significance of an Indicated Effect.

For the monthly series of imports to the U.S. from the European Economic Community, IOECD, from January, 1974 through December of 1984, the trading day regression F-statistic from the X-11-ARIMA table with (6,124) degrees of freedom has the value 6.0. This would be highly significant if the regression assumptions leading to the F-distribution were satisfied. However, X-I1-ARIMA uses an OLS regression on the estimated irregulars series, which is a correlated series resulting from a smoothing procedure, so a fundamental assumption is invalid. A well-fitting REG-ARIMA model with trading day regression variables accounts for correlation. We fit three such models to this data, each with a $(0,1,1)(0,1,1)_{12}$ ARIMA structure, and with the following regression variables:

(a) constant term, level shift in February, 1975 (model 1);
(b) constant term, level shift in February, 1975, trading day variables (model 2);
(c) constant term, level shift in February, 1975, trading day and leap year February variables (model 3).

The corresponding AIC values are $AIC_N^{(1)} = 2241.9$, $AIC_N^{(2)} = 2250.1$ and $AIC_N^{(3)} = 2252.0$, so model 1 is favored, contracting X-11-ARIMA's F-statistic. An alternative diagnostic, the smoothed periodogram of the irregulars series given in Fig. 3, has no peaks at the trading day frequencies, which supports the conclusion of the REG-ARIMA analyses: the series does not have a significant trading day component.

3.3 Detecting Quarterly Trading Day Effects.

It has long been assumed that, because the weekday composition of quarters is much less variable than that of calendar months, trading day effects would not be significant with quarterly economic series. However, we were sent some payroll series recently by Shelby Herman of the U.S. Bureau of Economic Analysis which had such effects, in her opinion. Our REG-ARIMA analyses confirmed her observations. For example, we fit three REG-ARIMA models to the logarithms of the payroll series NEM (Non-electrical Machine Manufactures from 1975/1 - 1988/4), with regression effects which included

(a) no quarterly trading day or leap year effects (model 1),
(b) quarterly trading day effects (model 2), and
(c) quarterly trading day and leap year effects (model 3).

The AIC values for the corresponding models are $AIC_N^{(1)} = 808.8$, $AIC_N^{(2)} = 786.4$, and $AIC_N^{(3)} = 785.0$.

Models 2 and 3 are both preferred over model 1, and their estimated trading day effects are almost identical. A graph of the trading day factors, which are antilogarithms multiplied by 100 of the trading day effects of model 3, is given in Fig. 4.

4. Comparing Subjective and REG-ARIMA Preadjustment Divisors

We have frequently been asked whether subjective preadjustment divisors (estimated by subject-matter experts) or model based preadjustment divisors are to be preferred. REG-ARIMA model comparisons offer an objective way to make such decisions on a case by case basis, as the following example illustrates. The unit auto sales series UAS of Fig. 5 has a number of extreme movements that are due to sales promotion campaigns by automobile manufacturers. These campaigns were used to reduce large dealer inventories by offering buyers low-interest loans or cash rebates. Such promotions increase car sales abnormally in months in which they are in effect and cause an atypical decrease in the following month or so. By analyzing the irregulars series from an X-11-ARIMA adjustment of UAS, an analyst obtained the adjustment divisors graphed in Fig. 6.

We were concerned that X-11-ARIMA seasonal adjustment, and therefore its estimate of the irregulars series, would be compromised by the fluctuations arising from the promotions. In this case, the promotion effects could not be obtained reliably from the irregulars.

It seemed better to us to use the outlier identification procedures of X-12-ARIMA (see Bell, 1983) together with some constraints suggested by the analyst to estimate the promotion effects. We fit such a REG-ARIMA model, along with trading day, fixed seasonal, and other additive outlier effects, to the logarithms of the observed series. For the resulting model (model 1), the estimated outlier effects are graphed in Fig. 7. The AIC value of this model is $AIC_N^{(1)} = 3169.4$. The logarithm series adjusted for the analyst's estimate (as in 4. of Sec. 1) was also fitted with a somewhat different REG-ARIMA model (model 2) whose regression variables included trading day, fixed seasonal and different additive outliers, which in some cases contradicted the analyst's adjustment. Even though no parameter estimation penalty was assigned for the analyst's estimates of the promotion effect

(we dind't know how, because the estimates were not obtained via maximum likelihood estimation), the AIC value for this model is much larger, $AIC_N^{(2)} = 3187.0$ We conclude that model 1 better describes the data, and therefore that the estimates of the promotion effects obtained via regression terms in the REG-ARIMA model are better than those obtained from examining the X-11-ARIMA irregulars series. Other analyses also support this conclusion.

REFERENCES

Bell, W.R. (1983). "A Computer Program for Detecting Outliers in Time Series." *Proc. Bus. Econ. Sec. ASA*, 634-639.

Bell, W.R. (1984). "Seasonal Decomposition of Deterministic Effects." Statistical Research Division Report No. CENSUS/SRD/RR-84/01. U.S. Bureau of the Census, Washington, D.C.

Bell, W.R. and Hillmer, S.C. (1983). "Modeling Time Series with Calendar Variation." *JASA* 78, 526-534.

Brockwell, P.J. and Davis, R.A. (1987). *Time Series: Theory and Methods*. New York: Springer Verlag.

Findley, D. F. (1988). "Comparing Not Necessarily Nested Models with the Minimum AIC and Maximum Kullback-Leibler Entropy Criteria: New Properties and Connections." *Proc. Bus. Econ. Sec. ASA*, 110-118.

Findley, D.F., Monsell, B.C., Otto, M.C. and Pugh, M.G. (1988). "Toward X-12-ARIMA." *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., pp. 591-624.

Findley, D.F., Monsell, B.C., Shulman, H.B., and Pugh, M.G. (1990). "Sliding Spans Diagnostics for Seasonal and Related Adjustments." *JASA* 85 (to appear).

Figure 1: NRS

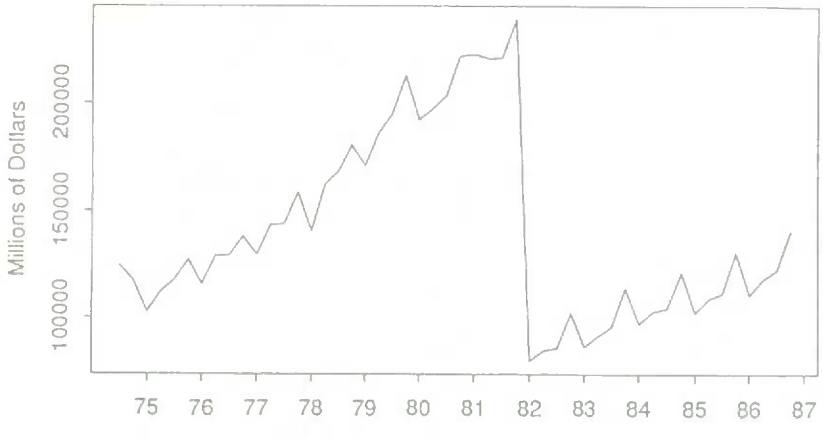


Figure 2: NWTAT

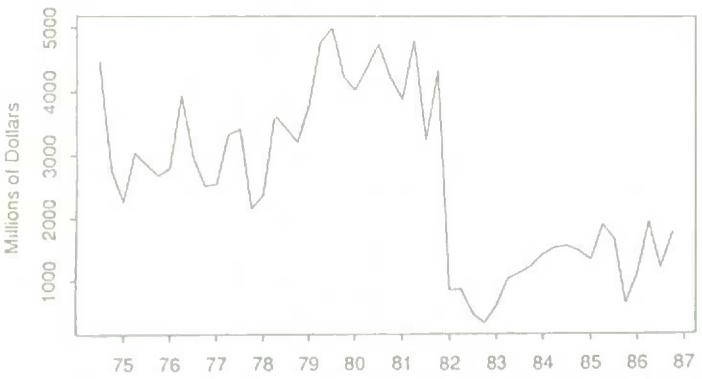


Figure 3: Smoothed Periodogram of the Irregulars from IOCDE

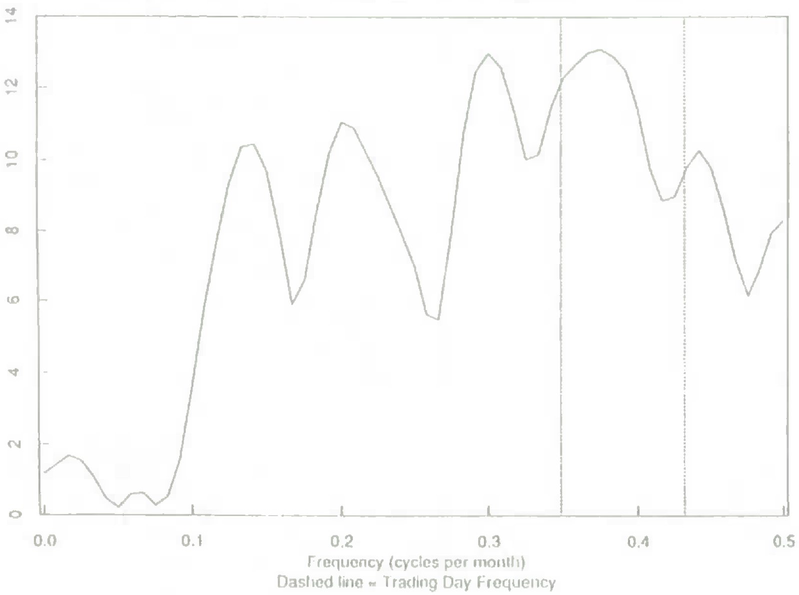


Figure 4: QUARTERLY TRADING DAY FACTORS FOR NEM

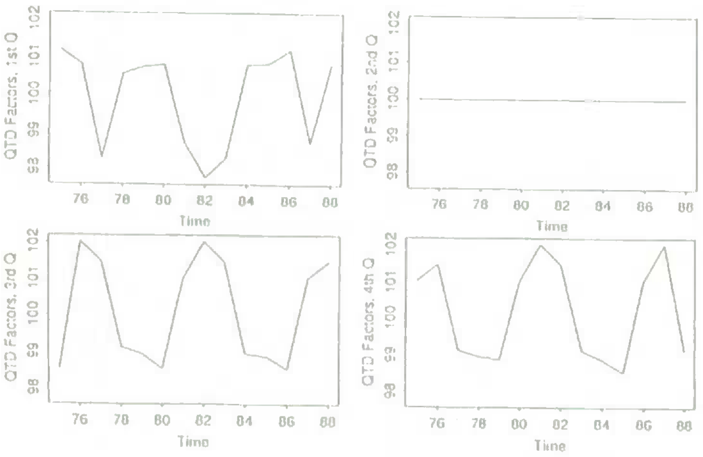


Figure 5: MONTHLY UNIT AUTO SALES

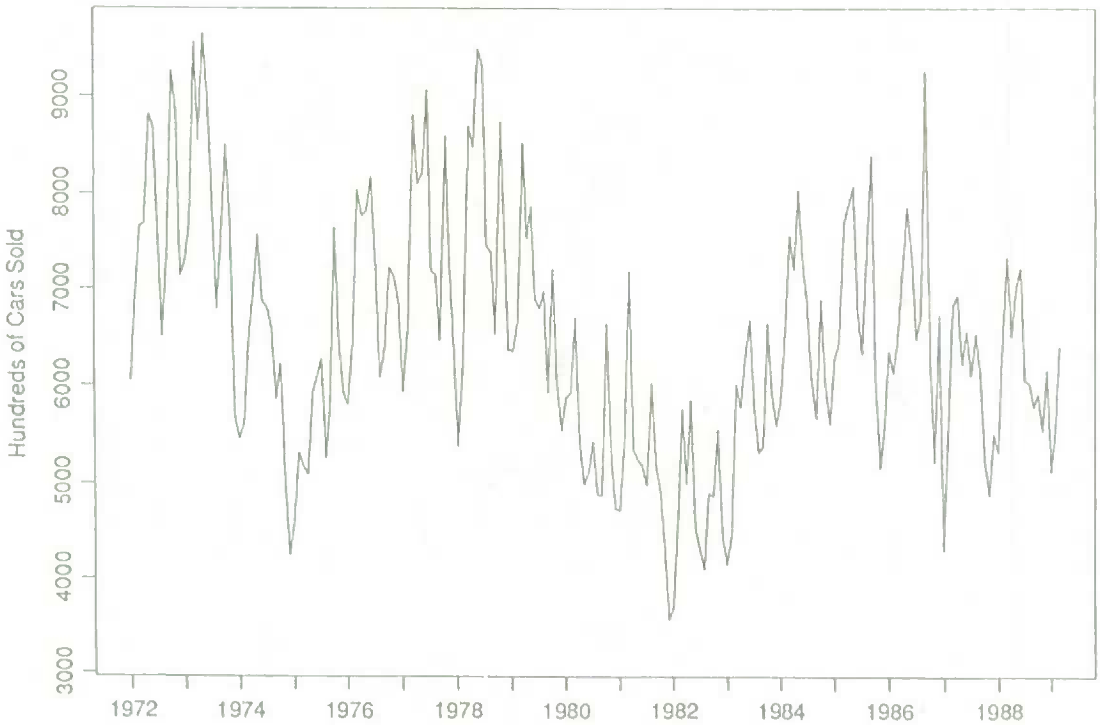


Figure 6: SUBJECTIVE PREADJUSTMENT DIVISORS FOR UNIT AUTO SALES

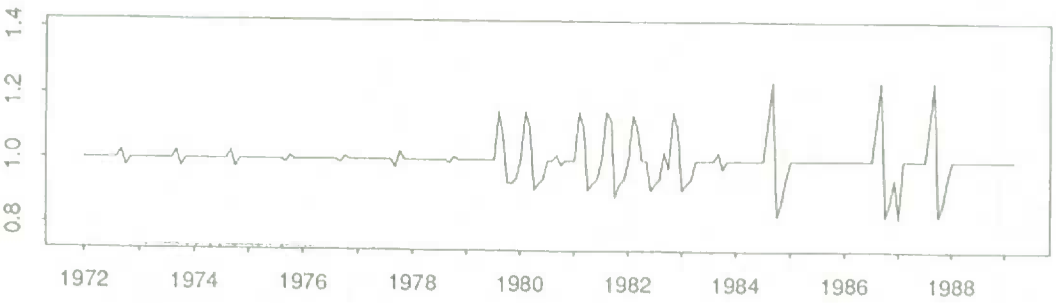
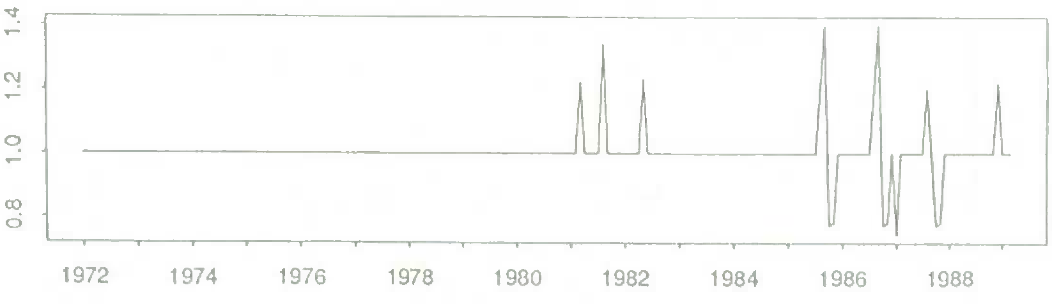


Figure 7: REG-ARIMA PREADJUSTMENT DIVISORS FOR UNIT AUTO SALES



Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

BENCHMARKING OF ECONOMIC TIME SERIES

Normand Laniel and Kimberley Fyfe[1]

ABSTRACT

KEY WORDS: Survey errors, non-linear model, weighted least squares, iterative proportional fitting.

Benchmarking is the improvement of estimates from a sub-annual survey with the help of corresponding estimates from an annual survey. For example, estimates of monthly retail sales might be improved using estimates from the annual survey. This article deals, first of all, with the problem posed by the benchmarking of time series produced by economic surveys, and then reviews the most relevant methods for solving this problem. Next, two new statistical methods are proposed, based on a non-linear model for sub-annual data. The benchmarked estimates are then collected by applying weighted least squares and the raking ratio method to maintain consistency among the tables in the series.

1. INTRODUCTION

Traditionally benchmarking has been defined as the problem of adjusting monthly or quarterly figures derived from one source to annual values (benchmarks) obtained via another source (see Denton 1971, Cholette 1988a, and Monsour and Trager 1979). For example, it could be the monthly shipments of Canadian Manufacturers which are adjusted so that they add up to the Annual Census of Manufacturers shipments figures. Another definition of benchmarking is the more general problem of improving sub-annual estimates derived from one source with annual estimates obtained via a second source (see Hillmer and Trabelsi, 1987). This definition assumes that the annual values are subject to error which is not the case with the first definition. For example, it could be the monthly inventories of Canadian Retailers derived from a sample survey which are improved in using the end of year inventories obtained from the annual retail trade sample survey. This second definition of the benchmarking problem corresponds to the situation encountered with most economic time series at Statistics Canada and it is the one dealt with in this paper.

The purpose of this article is fourfold. First, it formulates in detail, the benchmarking problem as it appears for most of the Statistics Canada time series produced by large scale economic surveys. Then, the most popular existing benchmarking methods dealing with a single time series are presented and discussed. Since all these methods fail in some respects to solve the Statistics Canada problem, two statistically based methods dealing with a single time series are proposed. These two methods use a non-linear weighted least squares approach. Finally, the benchmarking of a table of time series and preliminary benchmarking are discussed.

2. PROBLEM FORMULATION

The Statistics Canada problem of improving a table of sub-annual series of estimates with annual series of estimates from business surveys is formulated here, describing the characteristics of the original data and what is desired from a benchmarking procedure.

The sub-annual data is often biased due to frame coverage deficiencies. First, some new businesses have usually been in operation for a while before being included on the frame. This causes undercoverage. Another source of undercoverage is non-employer businesses (usually small) which are not represented on the sub-annual frame. The last coverage deficiency is the duplication which exists between the list of large businesses and the list of small businesses used as a frame by sub-annual surveys. Consequently, sub-annual estimated totals are more than likely biased. Another characteristic of the sub-annual data is that it is derived from overlapping samples. This implies that sampling covariances exist between sub-annual estimates of different time periods.

In regards to the annual data, in practice they can be assumed to be unbiased since they do not suffer much from duplication and the annual frame covers non-employer businesses and most new businesses. Also, the annual data usually come from large overlapping samples and thus have sampling errors associated with them.

When applying a benchmarking procedure it has to be taken into consideration that the results from the annual surveys come in approximately two years after the time that they are relevant. For example, annual data for 1988 will not be released until some time in 1990, while sub-annual data are usually available a few months after the time period that they are relevant. Therefore, when the sub-annual data are to be benchmarked, there will be no annual benchmarks for some of the sub-annual periods.

[1] Normand Laniel and Kimberley Fyfe, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

There are a number of features that a benchmarking procedure should have in order to be used for large scale survey estimates. First, the procedure should be simple enough so that it can be used in an automatic fashion without too much intervention from the statistician. Secondly, it must be possible to produce preliminary benchmarking factors for months for which benchmarks are not available yet. This characteristic allows benchmarking to be performed as the sub-annual data are produced. Otherwise discontinuities will be introduced in the sub-annual data.

The benchmarking method should be capable of improving the level estimates and the year-to-year trend estimates of either flow (i.e. data that refers to an interval of time such as sales) or stock (i.e. data that refers to a point in time such as inventory) sub-annual data. Another desirable characteristic is that the method maintains consistency between the table grand-totals, marginal totals, and cell estimates for the benchmarked data.

3. BENCHMARKING A SINGLE SERIES

The following sub-sections outline four potential approaches that one could use for benchmarking a single time series of sub-annual flow or stock data. Each approach is presented with a statistical interpretation, a brief outline of the underlying assumptions and a qualitative evaluation of the appropriateness to the problem detailed in section 2.

3.1 Denton's method

In his 1971 paper, Denton proposed procedures for a benchmarking approach based on Quadratic Minimization. Each corresponds to a specific penalty function. Of these, one could be applied to the problem of benchmarking time series as described in section 2, if some assumptions on the data are met. The procedure of interest uses a penalty function in terms of proportionate first differences between the original and benchmarked series. It can be presented in statistical terms by first assuming that the sub-annual data follows the model:

$$\frac{\theta_t}{y_t}=\frac{\theta_{t-1}}{y_{t-1}}+\epsilon_t \qquad t=1,2,\ldots,n$$

restricted to the annual data:

$$z_T=\sum_{t\in T}\theta_t \qquad T=1,2,\ldots,m$$

where:

- $\{y_t\}$ is a sequence of biased estimates of the sub-annual parameters (levels),
- $\{\theta_t\}$ is a sequence of fixed sub-annual parameters (true values of the levels),
- $\{\epsilon_t\}$ is a sequence of uncorrelated and identically distributed errors with mean vector and covariance matrix $(\underline{0},\sigma^2\underline{I})$ and,
- $\{z_T\}$ is a sequence of annual benchmarks obtained from a census.

To find the benchmarked estimates, least squares are applied to the above restricted model.

It is important to note that Denton's approach implies that θ_t/y_t follows a random walk and that the annual data is from a census. Unfortunately, these assumptions are unlikely to be satisfied by economic time series. Even though this method is able to handle a bias in the sub-annual data, it does not take into account the sampling variances and covariances of the sub-annual and annual data and, therefore, it is not statistically efficient.

3.2 Hillmer and Trabelsi's Method

In 1987, Hillmer and Trabelsi proposed an approach to the benchmarking problem based on the Box-Jenkins (1976) ARIMA models. They assumed that the sub-annual data follows the model:

$$y_t=\theta_t+\epsilon_t \qquad t=1,2,\ldots,n$$

and the annual data follows the model:

$$z_T=\sum_{t\in T}\theta_t+\alpha_T \qquad T=1,2,\ldots,m$$

where:

- $\{\theta_t\}$ is a sequence of stochastic sub-annual parameters (true values of levels) following an ARIMA model,
- $\{y_t\}$ is a sequence of unbiased estimates of the sub-annual parameters,
- $\{\epsilon_t\}$ is a sequence of sub-annual dependent sampling errors with mean vector and covariance matrix $(\underline{0},\underline{\Sigma}_\epsilon)$,
- $\{z_T\}$ is a sequence of annual unbiased estimates, and
- $\{\alpha_T\}$ is a sequence of annual dependent sampling errors with mean vector and covariance matrix $(\underline{0},\underline{\Sigma}_\alpha)$.

Using the above models, they obtain the benchmarked sub-annual estimates by applying stochastic least squares. That is, they minimize $E(\hat{\theta}_t-\theta_t)^2$, the mean squared error. This technique is also referred to in time series terminology as signal extraction, and the derivation of the solution can be found in the paper written by Hillmer and Trabelsi.

With this method, the annual data can come from either a census or a survey using overlapping samples. It also takes into account the sampling variances and covariances of the sub-annual level estimates. Unfortunately, the approach does not accommodate biases in the sub-annual data which is the case with economic surveys. Also, since ARIMA modelling is being used in this method, it would be costly to implement for large scale surveys dealing with hundreds of series. Therefore it would be best to use this type of approach for only a small number of very important economic indicators. There would also be risks of oversmoothing the data if the ARIMA models are not properly specified.

Cholette and Dagum(1989) improved upon the Hillmer and Trabelsi approach by using an "intervention" model instead of an ARIMA model. This allows the modelling of systematic effects in the time series but according to the authors, this improved approach still possesses the same weaknesses as the original Hillmer and Trabelsi method.

3.3 Model on Trends

The following method was developed in an attempt to meet the benchmarking requirements of the economic surveys. It is based on the assumption that the sub-annual data follows the model:

$$\frac{y_t}{y_{t-1}} = \frac{\theta_t}{\theta_{t-1}} + \epsilon_t \qquad t = 1,2,\ldots,n$$

and the annual data follows the model:

$$z_T = \sum_{t \in T} \theta_t + a_T \qquad T = 1,2,\ldots,m$$

where:

$\{y_t/y_{t-1}\}$ is a sequence of (nearly) unbiased estimates of the sub-annual trends,
$\{\theta_t/\theta_{t-1}\}$ is a sequence of trends of the fixed sub-annual parameters (true values),
$\{\epsilon_t\}$ is a sequence of dependent sub-annual sampling errors with mean vector and covariance matrix $(\underline{0}, \sum_\epsilon)$,
$\{z_T\}$ is a sequence of annual unbiased estimates, and
$\{a_T\}$ is a sequence of annual dependent sampling errors with mean vector and covariance matrix $(\underline{0}, \sum_a)$.

Least squares theory is applied to the above models to produce benchmarked estimates. The description of the Gauss-Newton algorithm necessary to solve this problem is given in the appendix and is followed by the calculation of the covariance matrix of the benchmarked estimates.

This method can be used when the benchmarks come from either a census or annual overlapping samples and when the sub-annual level estimates are biased, if the relative bias is a constant. The assumption of a constant relative bias will be verified in practice when the rate of the frame maintenance activities is relatively stable. That is, when the proportion of frame coverage deficiencies is fairly constant over time. Also the undercovered businesses have to behave like the ones covered by the frame. These assumptions would be verified if the benchmarking procedure were applied on a small number of years of data at a time.

There is one technical problem with this method. The sampling variance-covariance matrix of the trends cannot be calculated directly and an approximation has to be used. The first-order Taylor approximation has been tried but in some cases the resulting sampling variances and covariances were zero or negative when they should be positive.

3.4 Model on Levels

The following method is somewhat equivalent to the previous one and was developed so that the sampling variance-covariance matrix of the sub-annual estimates would be easier to obtain. It assumes that the sub-annual data follows the model:

$$y_t = \alpha\theta_t + \epsilon_t \qquad t = 1,2,..n$$

and the annual estimates follows the model:

$$z_T = \sum_{t \in T} \theta_t + a_T \qquad T = 1,2,\ldots,m$$

where:

$\{y_t\}$ is a sequence of biased estimates of the sub-annual levels,
α is a fixed parameter taking into account the constant relative bias,
$\{\theta_t\}$ is a sequence of fixed sub-annual parameters (true values of levels),
$\{\epsilon_t\}$ is a sequence of dependent sub-annual sampling errors with mean vector and covariance matrix $(\underline{0}, \sum_\epsilon)$,
$\{z_T\}$ is a sequence of unbiased annual estimates, and
$\{a_T\}$ is a sequence of dependent annual sampling errors with mean vector and covariance matrix $(\underline{0}, \sum_a)$.

Benchmarked estimates are found by applying least squares theory to the above models. The algorithm required to solve this problem is the same as for method 3.3.

This method can be used when the annual data come from either a census or from overlapping samples, and when the sub-annual data has biased level estimates if the relative bias is a constant over time.

3.5 Discussion

Amongst the methods reviewed here, the most appropriate one for benchmarking a single time series is the new approach based on the model on levels. It has a statistical basis which allows us to calculate confidence regions and test the goodness of fit of the benchmarked model. To test for lack of fit one has to be careful in choosing a test since the benchmarked estimates, θ_t, have quite a small number of degrees of freedom, $m-1$ (the number of annual observations minus one), in comparison to the number of observations, n+m. This also suggests that we can expect to get benchmarked estimates with a chronological pattern similar to the one observed in the sub-annual data.

At this point in time, the derivation of sampling covariances between two level estimates corresponding to two different time periods is a practical issue. Should they be directly calculated for all pairs of time periods with an estimation computer system or modelled? From a theoretical point of view, it is better to calculate these directly, since the sequence of sampling errors is intrinsically a non-stationnary stochastic process. However, it is not evident that this is feasible. On the other hand, no model exists which has been validated. In the literature, some authors have arbitrarily tried an AR(1) stationary model (see Hillmer and Trabelsi, 1987). This model does not look valid a priori. A slightly different approach has been attempted by Quenneville and Srinath (1984) by modelling the sampling correlations between time periods by the autocorrelation pattern of an AR(1) process. The validity of this last attempt is not clear. Thus, the question of obtaining sampling covariances is still open.

4. BENCHMARKING A TABLE OF TIME SERIES

Most economic sub-annual surveys produce series of estimates for a number of industrial activities within a number of geograghical regions. These are published sub-annually in the form of tables, where the cells as well as the marginals and the grand totals need to be benchmarked.

If one applies a benchmarking method independently on each cell series, each marginal series and the grand total series, the results will be a series of benchmarked sub-annual estimates where the sums of the cell totals are not equal to the marginal totals, and the sum of the marginal totals are not equal to the grand total. In other words, a series of inconsistent tables will be produced. To avoid this problem, a number of strategies can be adopted. Amongst these strategies, the first that comes to mind is the following simple approach. First, the cell series are independently benchmarked. Then, the benchmarked cell totals are summed up to get the benchmarked marginal totals and benchmarked grand totals. With this method one might get benchmarked margins and grand totals with chronological patterns which look more noisy than if they were directly benchmarked (this is a problem well known in seasonal adjustment). If this is the case one would be better to use the following method:

i) First benchmark the series of grand totals.
ii) Then, independently benchmark each series of marginal totals and then for each sub-annual period separately adjust the benchmarked margins by a constant factor so that they add up to the benchmarked grand totals.
iii) Finally, independently benchmark each series of cell totals and then for each sub-annual period separately adjust the benchmarked cells using the raking ratio algorithm (also called iterative proportional fitting, see Deming and Stephan, 1940) so that they add up to the adjusted benchmarked margins.

This method assumes that the series of grand totals is the most important series of the table in terms of preserving month-to-month trends, the series of marginal totals are the second most important and the series of cell totals are the least important. An inconvenient with this method is that the month-to-month trends of the cells can be very much disturbed. This has been observed in a small number of cases (see Laniel and Fyfe, 1989).

One can also think of benchmarking simultaneously the cell series with the margin series and the grand total series. Then the problem can become very large in terms of the number of parameters to estimate and even difficult to handle with a computer. This has been addressed by Cholette (1988b) in the case where series are to be benchmarked with Denton's method.

More evaluation and analysis needs to be done on these three possible approaches in order to determine which one should be used for the problem described in section 2.

5. PRELIMINARY BENCHMARKING

Preliminary benchmarking is performed to avoid discontinuities between the sub-annual periods with and without corresponding annual data. This is due to the fact that the annual data is available approximately 18 months after the end of the calendar year that it belongs to. Hence, there are two sets of sub-annual periods without corresponding annual data. The first contains periods for which sub-annual estimates are available. The second set consists of the periods for which sub-annual data will only be available at the time of the next application of the benchmarking procedure. This is assuming that benchmarking is an annual event. Therefore, when the benchmarking procedure is applied, it should produce projected benchmarking factors which can be used to give preliminary benchmarked data.

Two main approaches to produce preliminary benchmarking factors are:

1) Repeat the factor that was produced for the last benchmarked sub-annual period by either:
   a) benchmarking up to the last sub-annual period with corresponding annual data, or
   b) benchmarking up to the last sub-annual period with sub-annual data available.
2) Use a model to extrapolate the sub-annual series up to the sub-annual period where the next application of the benchmarking procedure is to occur. Then apply the benchmarking procedure using the extrapolated sub-annual series to get the preliminary factors. Simple models to do such extrapolations have been suggested by Laniel (1986). It should be verified that these models are robust enough for a large scale survey system not to provide preliminary factors which are less reliable than a procedure that simply repeats the last calculated benchmarking factor.

These two approaches should be investigated and evaluated. Such an evaluation might consist of looking at revisions in the benchmarked data from preliminary to final figures.

6. CONCLUSION

The problem of improving sub-annual survey estimates with annual survey estimates has been examined. A new and simple procedure to benchmark a single time series has been presented. This procedure could be implemented in a computer system which could be used in an automatic mode. The advantage of the procedure over more traditional methods is its statistical basis. Confidence regions can be derived and goodness of fit of the benchmarking model can be assessed. Some issues in using the proposed procedure for benchmarking a single time series have been discussed. Two major practical questions have been pointed out: benchmarking a table of series and preliminary benchmarking. Approaches to address these two topics have been suggested but more work remains to be done.

7. REFERENCES

Box, G.E.P. and Jenkins, G.M. (1976), "Time Series Analysis, Forecasting and Control", Holden-Day.

Cholette, P.A. (1988a), "Benchmarking and Interpolation of Time Series", Statistics Canada, Working Paper No. TSRA-87-014E.

Cholette, P.A. (1988b), "Benchmarking Systems of Socio-Economic Time Series", Statistics Canada, Working Paper No. TSRA-88-017E.

Cholette, P.A. and Dagum, E.B. (1989), "Benchmarking Socio-Economic Time Series Data: A Unified Approach", Working Paper No. TSRA-89-006E, Statistics Canada.

Deming, W.E. and Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known", Ann. Math. Statist.

Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An approach Based on Quadratic Minimization", Journal of the American Statistical Association, Vol. 66, No 333, pp. 99-102.

Hillmer, S.C. and Trabelsi, A. (1987), "Benchmarking of Economic Time Series", Journal of the American Statistical Association, Vol. 82, pp. 1604-1071.

Laniel, N. (1986), "Adjustment of Economic Production Sub-annual Series", Business Survey Methods, Statistics Canada, Working Paper No. BSMD-86-006E.

Laniel, N. and Fyfe, K. (1989), "Benchmarking of Economic Time Series", Business Survey Methods, Statistics Canada, Business Survey Redesign Project Working Paper.

Monsour, N.J. and Trager, M.L. (1979), "Revision and benchmarking of Business Time Series", Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Quenneville, B. and Srinath, K.P. (1984), "Estimation of Variances of Averages Based on Overlapping Samples in Repeated Surveys", Proceedings of the Survey Research Methods Section, American Statistical Association.

APPENDIX

1. GAUSS-NEWTON ALGORITHM

The models for the sub-annual and annual estimates of sub-sections 3.3 or 3.4 can be cast into one model of the form:

$$Y_s = f(\underline{X}_s, \underline{\gamma}) + u_s \qquad for \quad s = 1, \dots, n+m$$

where: Y_s represents the sub-annual response when s=1,...,n and the annual response when s=n+1,...,n+m,

$\underline{X}_s$ equal to $(X_{1s}, \dots, X_{n+m,s})'$ is a vector of dichotomic variables defined as:

$$X_{ks} = \begin{cases} 1 & if \quad s = k \\ 0 & if \quad s \neq k \end{cases} ;$$

$\underline{\gamma}$ equal to $(\gamma_1, \dots, \gamma_p)'$ is the vector of parameters to be estimated in the combined sub-annual and annual model,

u_s is the sub-annual sampling error when s=1,...,n and the annual sampling error when s=n+1,...,n+m; and

$f(\underline{X}_s, \underline{\gamma})$ is equal to $\sum_{k=1}^{g} g_k(\underline{\gamma}) X_{ks}$ with $g_k(\gamma)$ representing the sub-annual model when k=1,...,n and the annual model when k=n+1,...,n+m.

For example, in the case of approach 3.4, we have $\underline{\gamma} = (\alpha, \theta_1, \dots, \theta_n)'$ and

$$g_k(\underline{\gamma}) = \begin{cases} \alpha\theta_k & if \quad k = 1, \dots, n \\ \sum_{i \in k} \theta_i & if \quad k = n+1, \dots, n+m. \end{cases}$$

Both sub-annual models in 3.3 and 3.4 are non-linear in the parameters. The linearization method can be used in such a case to estimate the parameters which consists of approximating the non-linear model by a linear one of the form $Y_s - f_s^* = \sum_{i=1}^{p} \beta_i^* J_{is}^* + u_s$

where $f_s^* = f(\underline{X}_s, \underline{\gamma}_o)$, $\beta_i^* = \gamma_i - \gamma_{io}$, $\underline{\gamma}_o = (\gamma_{1o}, \dots, \gamma_{po})'$, and $J_{is}^* = \left[\dfrac{\partial f(\underline{X}_s, \underline{\gamma})}{\partial \gamma_i}\right]_{\gamma = \gamma_o}$.

are initial estimates close to the true values. In our benchmarking application we have used the Denton method (see 3.1) to get these initial values.

The initial estimates are improved by using linear least squares in successive iterations, which leads to the following updating matrix equation:

$$\underline{\gamma}_{j+1} = \underline{\gamma}_j + \underline{L}_j'(\underline{Y} - \underline{f}^j)$$

where: $\underline{\gamma}_j = (\gamma_{1j}, \dots, \gamma_{pj})'$ $\quad \underline{f}^j = (f_1^j, \dots, f_{n+m}^j)'$ $\quad \underline{J}_j = (J_{is}^j)_{(n+m) \times p}$ $\quad \underline{u} = (u_1, \dots, u_{n+m})'$

$\underline{Y} = (Y_1, \dots, Y_{n+m})'$ $\quad \underline{L}_j = \left[\left(\underline{J}_j' \underline{\Sigma}_u^{-1} \underline{J}_j\right)^{-1} \underline{J}_j' \underline{\Sigma}_u^{-1}\right]'$ $\quad \underline{\Sigma}_u = E(\underline{u}\underline{u}')$

For this benchmarking application, computer rounding errors may cause the matrix $\underline{J}_j' \underline{\Sigma}_u^{-1} \underline{J}_j$ to look singular and thus non-invertible. This is due to a large difference in the size of some of the elements of $\underline{J}_j$ and can be overcome by simply dividing both the sub-annual and annual series by the average of the sub-annual levels before using the iterative algorithm. Once it has converged, the sub-annual benchmarked estimates are then obtained by multiplying back with that average.

The above iterative process has many convergence problems which are well described with solutions in Draper and Smith (1981). In the case of approach 3.4 these problems can be reduced by exploiting the structure of the model. One can use the following two step procedure: i) for fixed α, get linear weighted least squares estimators of the θ_i's as functions of α, say $\theta_i(\alpha)$, and ii) then use $\hat{\theta}_i(\alpha)$ in place of θ_i in the benchmarking model and apply nonlinear WLS to get an estimate of α, say $\hat{\alpha}$. This way, the dimension of the Gauss-Newton algorithm is reduced from $n+1$ to 1.

The expression above for $\underline{\gamma}_{j+1}$ assumes that annual values are observed with errors so that $\underline{\Sigma}_u$ is non-singular. However this covariance matrix will be singular, when the annual values come from a census. In such a case, the solution can be obtained with the minimum $\underline{\Sigma}_u$-seminorm g-inverse replacing $\underline{L}_j$. That is, $\underline{\Sigma}_u$ is replaced by $\underline{\Sigma}_u + \underline{J}_j \underline{J}_j'$ in the equation for $\underline{L}_j$ (see Rao and Mitra, 1971).

2. VARIANCE-COVARIANCE MATRIX FOR THE ESTIMATES

Assuming that the Gauss-Newton algorithm has converged after j iterations to the estimates $\hat{\underline{\gamma}} = \underline{\gamma}_j$, then the approximated covariance matrix is given by $Var(\hat{\underline{\gamma}}) = \underline{L}_j' \underline{\Sigma}_u \underline{L}_j$.

3. MORE REFERENCES

Draper, N.R. and Smith, H. (1981), Applied Regression Analysis", Second Edition, New York: Wiley.

Rao, C.R. and Mitra, S.K. (1971), "Generalized Inverse of Matrices and Its Applications", New York: Wiley.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

TRANSFORMING FISCAL QUARTER DATA INTO CALENDAR QUARTER VALUES

P.A. Cholette[1]

ABSTRACT

Many quarterly surveys carried out by statistical agencies reflect the fiscal quarters of the respondents, covering for instance the months from February to April, May to July, etc. This paper proposes a method to transform such data into calendar quarter estimates, covering from January to March, April to June, etc.

The method is essentially an adaptation of the Denton (1971) benchmarking method: A monthly seasonal pattern is benchmarked to be consistent with the available fiscal quarter benchmarks. The calendar quarter estimates are then simply the appropriate calendar quarter sums of the monthly "benchmarked" values. The Denton method is presented anew in the familiar framework of regression analysis.

KEY WORDS: Benchmarking, Interpolation, Fiscal Quarters, Fiscal Years, Temporal Disaggregation.

1. INTRODUCTION

All the quarterly surveys conducted by Statistics Canada actually refer to the financial, i.e. fiscal, quarters of the respondents. These quarters cover any of three consecutive months: for example, February to April, May to July, etc.; or March to May, June to August, etc. Sometimes those "months" do not even end on the last day of months. In some cases of course, the fiscal quarters coincide with the calendar quarters, covering from January to March, April to June, etc.

One practice with respect to fiscal quarter data, is to assign them to the calendar quarter, which overlaps the most. For instance if the respondents to a survey have any one of the following fiscal quarters, December to February, January to March and February to April, their responses are all assigned to the first quarter. The "quarterly" total of those responses thus implicitly covers five months (December to April), instead of the first quarter. In a seasonal situation especially, such quarterly values are obviously misleading.

This paper proposes a method to calendarize fiscal quarter data, that is to transform them into calendar quarter values. It is assumed (1) that the respondents in the survey have common fiscal quarters, or at least that calendarization is performed at a level where this is the case; and (2) that the fiscal quarters end at the end of months. Section 2 illustrates the calendarization problem under those simplifying assumptions.

Section 3 presents the additive variant of the proposed calendarization method, which is in fact an adaptation of the benchmarking methods of the Denton type (e.g. Denton, 1971; Helfand, Monsour and Trager, 1977). (Benchmarking consists of adjusting a sub-annual series to annual values obtained from another more reliable source.) Section 4 introduces a logarithmic variant of the proposed method. Section 5 suggests an economical implementation of both variants and examines the issue of revising the estimates. Section 6 tests the method on ten Canadian retail trade series.

2. THE CALENDARIZATION PROBLEM

The problem of calendarizing fiscal quarter data is easily described by means of an illustration. Figure 1 displays three years of monthly sales by the Canadian Department Stores. The figure also displays the calendar quarter values, averaged over the three months they cover (i.e. divided by 3), and the fiscal quarter values (also divided by 3), which cover the months of February to April, May to July, etc. Taking the fiscal quarter values as an approximations for the closest calendar quarter (which overlaps the most) entails large and obvious "estimation" errors, especially for the first and the fourth calendar quarters of each year in the example. Furthermore, the errors constitute bias: every year, the first and the third quarters are *systematically* over-estimated while the second and the fourth quarters are under-estimated. In the presence of seasonality especially, considering fiscal quarters as calendar quarters, that is ignoring the calendarization problem, causes error and bias in the resulting quarterly series.

[1] Statistics Canada, Time Series Research and Analysis Division, Ottawa, Canada K1A 0T6.

In a true calendarization situation, the monthly values of Figure 1 are obviously unknown. The strategy proposed in this paper consists of two steps:

(1) interpolate the unknown monthly figures, from the fiscal quarter data and from an auxiliary variable, usually in the form of a seasonal pattern; and

(2) set the calendar quarter estimates equal to the calendar quarter sums of the monthly interpolations.

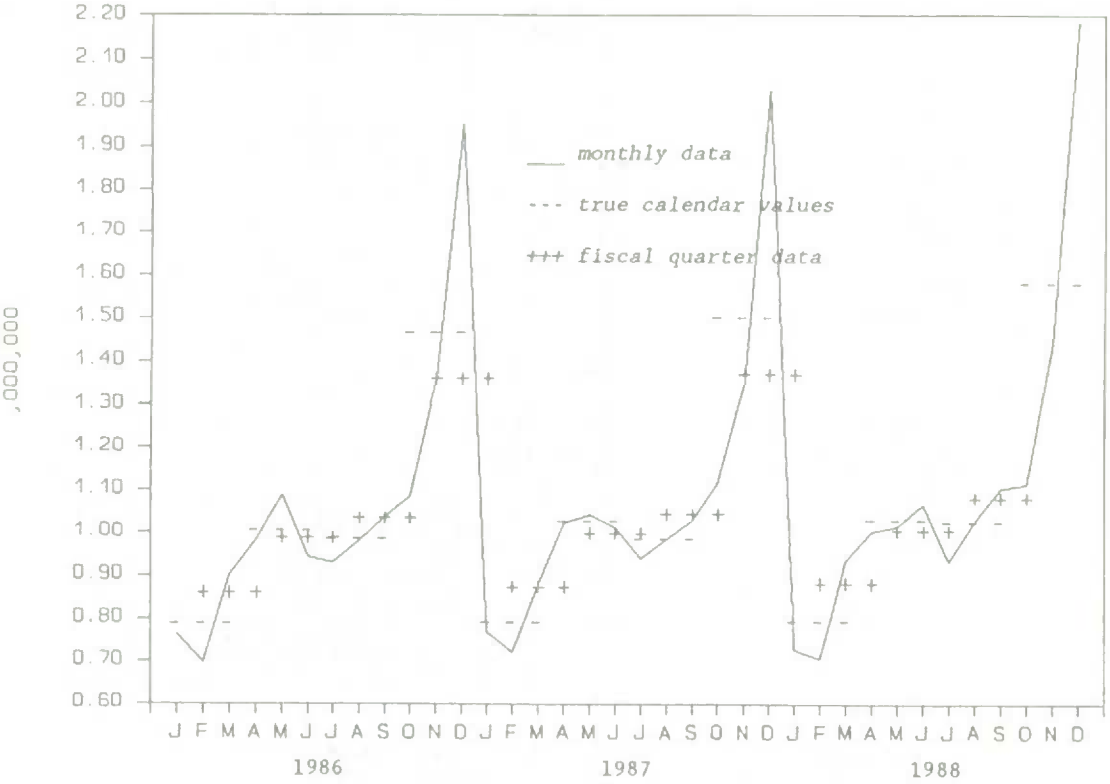


Figure 1: Differences between the calendar quarter values (---) and the fiscal quarter values (+++) of the Canadian monthly sales by Department Stores

3. THE ADDITIVE VARIANT OF CALENDARIZATION

This section presents the benchmarking methods of the Denton type as a linear regression model and adapts it for calendarization purposes. Statistical agencies normally use benchmarking when, for a socio-economic variable, sub-annual (say) measurements co-exist with annual measurements, obtained from an other more reliable source and considered as benchmarks. In such cases, the annual sums of the sub-annual series generally differ from the corresponding annual benchmarks. In the Denton-type methods, benchmarking then consists of adjusting the sub-annual series, so that (a) the annual sums of the benchmarked series conform to benchmarks and (b) the benchmarked series is as parallel as possible to the original sub-annual series. The calendarization method proposed basically consists of

(1) benchmarking a monthly seasonal pattern to the available fiscal quarter data, considered as benchmarks; and

(2) of taking the calendar quarter sums of the benchmarked series.

The first step produces the estimates of the unknown monthly values, i.e. the interpolations; and the second, the desired calendar quarter values.

3.1 The Model

As shown in Cholette and Dagum (1989), the Denton method can be seen as a regression containing two equations:

$$S = \Gamma + e, \quad E(e)=0, \quad E(e\ e') = V_e \ ; \qquad (3.1a)$$

$$F = B\ \Gamma + \epsilon, \quad E(\epsilon)=0, \quad E(\epsilon\ \epsilon') = V_\epsilon = \sigma_\epsilon^2\ I, \quad \epsilon \to 0, \quad \sigma_\epsilon^2 \to 0. \qquad (3.1b)$$

In the context of fiscal quarters, vector S of dimensions T by 1 stands for a monthly auxiliary variable. In this paper, and without loss of generality, S takes the form of a seasonal pattern or of a seasonal pattern plus a trading-day pattern (Young, 1965). This pattern is valid for all the respondents at the level at which calendarization is performed. Vector F of dimension M by 1 contains the fiscal quarter benchmarks, i.e. the data to be calendarized. Vector Γ contains the T unknown monthly values to be estimated.

Matrix B of dimensions M by T is a fiscal quarter sum operator. For example in the case of a flow series with fiscal quarters covering from February to April, May to July, etc, matrix B would be as follows:

$$\underset{\text{M by T}}{B} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & \end{bmatrix}, \qquad (3.2)$$

(For stock series, the two first 1's of each line are replaced by 0's.) Consequently, equation (3.1b) specifies that the fiscal sums of the desired interpolated values Γ are equal to the available fiscal quarter data (except for an infinitesimally small error whose presence will soon become obvious).

Finally the covariance matrix V_e of the disturbances $e = [e_t, t=1,...,T]$ is such that e_t changes as little as possible from month t to month t+1:

$$\underset{\text{T by T}}{V_e} = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix} \sigma_u^2 / (1-\rho^2). \qquad (3.3)$$

where ρ is lower but very close to 1 (0.999999) and where σ_u^2 is in practice the variance of change in S (i.e. the variance of (S_t-S_{t-1})). In other words, this matrix specifies that the disturbances are most autocorrelated at lag 1. (Details in Cholette and Dagum, 1989; Cholette and Baldwin, 1989). The effect of V_e in (3.1a) is to maintain the estimated interpolated values Γ^* as parallel as possible to the chosen seasonal pattern S. The degree of parallelism achieved depends on the fiscal quarter benchmarks in (3.1b).

3.2 The Solution

Model (3.1) can be written

$$Y = X\,\Gamma + U, \quad E(U)=0, \quad E(U\,U')=V, \qquad (3.4)$$

where:

$$Y' = [\;S' \;\; F'\;], \quad X' = [\;I \;\; B'\;], \quad U' = [\;e' \;\; \epsilon'\;], \quad V = \begin{bmatrix} V_e & 0 \\ 0 & V_\epsilon \end{bmatrix}.$$

The General Least Squares solution to (3.4) is:

$$\Gamma^\star = (X'V^{-1}X)^{-1}\; X'V^{-1}\; Y = [V_e^{-1} + B'V_\epsilon^{-1} B\,]^{-1}\; [V_e^{-1} S + B'V_\epsilon^{-1}\; F\,], \qquad (3.5)$$

$$\text{var } \Gamma^\star = (X'V^{-1}X)^{-1} = [V_e^{-1} + B'V_\epsilon^{-1} B\,]^{-1}, \qquad (3.6)$$

where V_e^{-1} is known algebraically:

$$V_e^{-1} = \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 & \ldots \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & \ldots \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} / \sigma_u^2. \qquad (3.7)$$

Using matrix algebra identities, solution (3.5) can be expressed:

$$\Gamma^* = S + V_e\, B'[B\, V_e\, B' + V_\epsilon]^{-1}\; [\,F-B\,S\,] = S + W\,[\,F-B\,S\,] = S + W\,R. \qquad (3.8)$$

Solution (3.8) requires a much smaller matrix inversion than (3.5). However, solution (3.8) would not be relevant if the variance (3.6) is calculated, which requires the larger matrix inversion.

Contrary to (3.5), solution (3.8) also allows $V_\epsilon=0$. Unless otherwise indicated, the rest of this paper will assume $V_\epsilon=0$. When $V_\epsilon=0$, the value of σ_u^2 implicit in V_e becomes immaterial, because it cancels out; and $1-\rho^2$ also cancels. Furthermore with $V_\epsilon=0$, (3.8) has the form of Denton's (1971) solution; and (3.8) is also the solution to minimizing the following constrained objective function:

$$\sum_{t=2}^{T} ((\Gamma_t-S_t) - \rho(\Gamma_{t-1}-S_{t-1}))^2 \quad - 2 \sum_{m=1}^{M} \lambda_m [(\sum_{\tau\in m} \Gamma_\tau) - F_m], \qquad (\Gamma_t-S_t=e_t).$$

As pointed by Bournay and Laroque (1979) for benchmarking, as ρ tending to 1, this function tends to that minimized by Denton and others (except for the constraints) and specifies that Γ preserves the month-to-month change observed on S.

3.3 The Calendarized Values

Whether the interpolations are obtained by (3.8) or (3.5), the desired calendar quarter estimates are simply the appropriate sums of Γ^*:

$$C^* = G \ \Gamma^*, \qquad G = I_N \otimes [1\ 1\ 1]. \tag{3.9}$$

where N is the number of calendar quarters and where I_N is the N by N identity matrix. The variance of C^* is obtained from that of Γ^*:

$$\mathrm{var}(C^*) = G\ \mathrm{var}(\Gamma^*)\ G'. \tag{3.10}$$

If one is not interested in the monthly interpolations per se, the calendar quarter estimates may be expressed directly in terms of the basic data F and S, by substituting (3.8) into (3.9):

$$C^* = G\ (S + W\ [\ F - B\ S\ ]) = G\ S + P\ [\ F - B\ S\ ]. \tag{3.11}$$

The weights W of (3.8) and P of (3.11) do not depend on the data F and S. They depend only on the length T of the series and on the **fiscal quarter pattern** considered, that is on whether the quarters end in January, April, July, etc., or in February, May, August, etc. The weights may then be considered as known in advance and be applied to any series with same length and fiscal pattern. As explained in section 5, this will entail important advantages for the implementation of the method.

4. THE LOGARITHMIC VARIANT

The additive method presented in section 3 is suitable when the seasonal-trading-day pattern S is of the same order of magnitude as the fiscal quarter data (divided by 3). However, S is more easily - and usually - expressed in percentages (in which case S is the product of a seasonal pattern and of a trading-day pattern). The additive variant would then yield interpolations Γ^* with negligible monthly sub-quarterly seasonality, in cases where the fiscal data are in millions (say). Such interpolations would generally be insufficiently accurate to produce a satisfactory calendarization of the fiscal quarters F.

Three options would then be available. One option would consist of multiplying the seasonal pattern by calibration factors which evolves gradually from month to month and then of applying the additive variant to the calibrated S. The second option would be to adapt the proportional variant of the Denton (1971) for fiscal quarters. Such a proportional variant would indeed solve the calibration problem by keeping Γ^* proportional to S; however, the weights W and P (of (3.8) and (3.11) would then depend on the data. The third option is to adopt the logarithmic variant now presented.

For stock series, the logarithmic variant merely consists of applying the additive variant to the logarithms of the fiscal quarter values, ln F, and of the seasonal pattern, ln S; and of setting the desired interpolations equal to the antilogarithm of the resulting estimates. Solution (3.8) thus becomes:

$$\ln \Gamma^* = \ln S + W\ [\ \ln F - B\ \ln S\ ], \qquad \Gamma^* = \exp(\ln \Gamma^*) \tag{4.1}$$

where the weights W are those of (3.8) (with $V_\epsilon=0$). Since the additive variant preserves the month-to-month change of S, Γ^* of (4.1) preserves the month-to-month growth rate.

For flow series, solution (4.1) also preserves growth rates, but the interpolations have their fiscal quarter products equal to the fiscal quarter benchmarks. In order to achieve equality of sums, one successful strategy is to iterate on ln F. Excellent starting values for $\ln F^{(1)}$ originate from the fiscal quarter products of S multiplied by the proportional discrepancies between F and S (in square brackets):

$$F_m^{(1)} = \prod_{t\in m} S_t\ [F_m / (\sum_{\tau\in m} S_\tau)] \quad \Rightarrow \quad \ln F_m^{(1)} = \sum_{t\in m} \ln S_t\ [F_m / (\sum_{\tau\in m} S_\tau)], \quad m=1,\ldots,M. \tag{4.2}$$

The first interpolations $\Gamma^{*(1)}$ are given by (4.1) applied to $\ln F^{(1)}$ of (4.2). For the other iterations ($k>1$), the revised values for $\ln F^{(k)}$, originate from the product of $F^{(k-1)}$ and the residual proportional discrepancies between F and $\Gamma^{*(k-1)}$ (in square brackets):

$$F_m^{(k)} = F_m^{(k-1)} [F_m /(\sum_{\tau\in m} \Gamma^*_\tau{}^{(k-1)})] \Rightarrow \ln F_m^{(k)} = \ln F_m^{(k-1)} + \ln [F_m /(\sum_{\tau\in m} \Gamma^*_\tau{}^{(k-1)})], \; m=1,\ldots M. \quad (4.3)$$

The subsequent interpolations $\Gamma^{*(k)}$ ($k>1$) are given by (4.1) applied to $\ln F^{(k)}$ of (4.3). Iteration between (4.3) and (4.1) takes place until the equalities (3.1b) (with $\epsilon=0$) are satisfied by more than 0.25% (say), which usually requires less then 5 iterations ($K\leq5$). An exact compliance to (3.1b) may be obtained by multiplying the last interpolations $\bar{\Gamma}^{*(K)}$ by the last residual proportional discrepancies:

$$\Gamma^*_t = \Gamma^*_t{}^{(K)} \; [F_m /(\sum_{t\in m} \Gamma^*_t{}^{(K)})] \quad (4.4)$$

The advantages of the logarithmic variant are the following:

(1) S may have an order of magnitude different from that of F; and

(2) the weights W of (3.8) are calculated once and for all and may be applied to any data S and F, regardless of their particular values.

These properties combine the advantages of the proportional and of the additive variants of Denton-type benchmarking.

5. IMPLEMENTATION

In both the additive and the logarithmic variants, the monthly interpolations are equal to the monthly seasonal (and trading-day) pattern chosen, plus a linear combination of the discrepancies R=F-BS between the fiscal quarter benchmarks and the corresponding sums of the seasonal pattern:

$$\Gamma^* = S + W [F - B S] = S + W R \Rightarrow \Gamma^*_t = S_t + \sum_{m=1}^{M} W_{t,m} R_m. \quad (5.1)$$

$$\ln \Gamma^{*(k)} = \ln S + W [\ln F^{(k)} - B \ln S] = \ln S + W R^{(k)}$$
$$\Rightarrow \ln \Gamma^*_t{}^{(k)} = \ln S_t + \sum_{m=1}^{M} W_{t,m} R^{(k)}_m. \quad (5.2)$$

where (5.1) is applicable to the additive variant and (5.2) to the logarithmic variant, and where the weights $W_{t,m}$ are given by (3.8) (with $V_\epsilon=0$).

5.1 Description of the weights

Table 1 contains the weights $W_{t,m}$ for the three possible regular fiscal quarter patterns, where each fiscal quarter contains three months. (Occasionally, a fiscal quarter contains more or less than 3 months.) Table 1A contains the weights to be applied when one fiscal quarter is available, i.e. M=1; table 1B, when two fiscal quarters are available, M=2; table 1C, M=3; and table 1D, M=4.

Column (a) of each sub-table pertains to the fiscal pattern where the quarters cover from February to April, May to July, etc. (or May to July, August to October, etc.); column (b), to the pattern where the quarters cover from March to May, June to August, etc.; and column (c) where the quarters are calendar quarters. The weights in column (c) may be used to interpolate monthly values from calendar quarters. For lack of space, column (c) is omitted in sub-tables C and D. However one can easily construct it from columns (b): row t of column (c) repeats row t-1 of column (b); in fact the same relation prevails between the rows of columns (b) and (a). One consequence of that relationship is that only columns (a) need be stored in practice.

5.2 Application when $M \geq 4$

When possible, we recommend applying the weights as a moving average of 4 fiscal quarters (M=4) embedded in 5 calendar quarters. For the sake of illustration, assume a series starting in 1986, which initially comprises 4 fiscal quarters covering from February to April, May to July, etc., embedded in 5 calendar quarters. As data become available, the series eventually comprises 11 fiscal quarters embedded in 12 calendar quarters. The weights of Table 1D (a) are then applied 8 times in the following manner.

The first time the weights apply to the 4 first fiscal data points F_m ($m=1,\ldots,4$) and to 15 seasonal data points S_t ($t=1,\ldots,15$) in the interval January 86 to March 87. For instance the additive interpolations for March and April 86 are

$$\Gamma^*_3 = S_3 + 0.35173\ R_1 - 0.02245\ R_2 + 0.00491\ R_3 - 0.00085\ R_4,$$

$$\Gamma^*_4 = S_4 + 0.24135\ R_1 + 0.11225\ R_2 - 0.02453\ R_3 + 0.00427\ R_4.$$

where the R_m's are the discrepancies of (5.1) between F_m and S_t. This produces the final interpolations for January to September 86 (and for the corresponding calendar quarters) and preliminary estimates for January to March 87.

The second time, i.e. when the 5th fiscal quarter value becomes available, the same weights apply to the data F and S in the interval April 86 to June 87. This produces the final estimates for October to December 86 (and for the corresponding calendar quarters), the revised estimates for January to March 87 and the preliminary estimates for April to June 87.

The third time, the weights apply to the data F and S in the interval July 86 to September 87. This produces the final estimates for January to March 87, the revised estimates for April to June 87 and the preliminary estimates for July to September 87.

And so forth. This application of the weights, over moving intervals of five calendar quarters, reduces the number of revisions to two; and insures that each final calendar estimate has two fiscal benchmarks "on each side" and is thus central in each interval. This implementation is much more economical than applying the method (recomputing W) on all the available data, and does not noticeably affect the estimates.

5.3 Reliability of Preliminary Estimates

As defined, the preliminary estimates are subject to higher revisions than the (once) revised estimates, because some of the months involved lie outside the range covered by the fiscal benchmarks. Under the fiscal pattern of column (a) the two last months lie outside; under pattern (b), the last month; and under pattern (c), the last three months. We therefore recommend that the preliminary estimates not be used, especially if a turning-point (downwards or upwards) in the business cycle is anticipated.

Statisticians not willing to tolerate the resulting production delays (or the reduced reliability) could supply a forecast of the next fiscal quarter benchmark and apply the method to the artificially extended series. (This tends to improve results with some seasonal adjustment methods, Dagum, 1980.) A good starting point of such a forecast is $F^f_m = F_{m-1} + F_{m-4} - F_{m-5}$. This forecast is that of a degenerate ARIMA model (0,1,0)(0,1,0). This model states that the change from one quarter to the next tends to repeat from one year to the next, which implies constant seasonality and *linear* trend-cycle over the last four quarters.

5.4 Application When M < 4

When only one fiscal quarter is available, covering from February to April 86 say, the weights of Table 1A (a) are applied. These weights simply distribute 1/3 of the discrepancy R_1 over the seasonal pattern S. The interpolations for the months of January to June 86 are then perfectly parallel to S, and contain only seasonality (unless S is not only seasonal). One can show that the calendarized values C_n ($n=1,2$) for the first and second quarters of 86 are trivial and equal to F_1. Calendarization is then unlikely to succeed when only one fiscal value is available.

When the second fiscal quarter of 86 becomes available, the weights of Table 1B (a) are applied to the data in the interval January to September 86. The interpolations are better than they were, and non-trivial calendarized values C_n ($n=1,2,3$) are now obtained for the first three quarters of 86. The first and second quarter trivial estimates obtained previously are revised. Similarly, when a third fiscal quarter of 86 becomes available, the weights of Table 1C are applied to the data in the interval January to December 86. All estimates previously obtained are revised.

Table 1

Weights $W_{t,m}$ applied to the quarterly discrepancies R_m to obtain the monthly interpolations under regular fiscal quarter patterns of columns (a), (b) and (c)

Table 1A: when only one fiscal quarter is available (M=1)

fiscal pattern:	(a)	(b)	(c)
t\m	F M A	M A M	J F M
J	0.33333	0.33333	0.33333
F	0.33333	0.33333	0.33333
M	0.33333	0.33333	0.33333
A	0.33333	0.33333	0.33333
M	0.33333	0.33333	0.33333
J	0.33333	0.33333	0.33333

Table 1B: when two fiscal quarters are available (M=2)

	fiscal pattern (a)		fiscal pattern (b)		fiscal pattern (c)	
t\m	F M A	M J J	M A M	J J A	J F M	A M J
J	0.40351	-0.07018	0.40351	-0.07018	0.40351	-0.07018
F	0.40351	-0.07018	0.40351	-0.07018	0.35088	-0.01754
M	0.35088	-0.01754	0.40351	-0.07018	0.24561	0.08772
A	0.24561	0.08772	0.35088	-0.01754	0.08772	0.24561
M	0.08772	0.24561	0.24561	0.08772	-0.01754	0.35088
J	-0.01754	0.35088	0.08772	0.24561	-0.07018	0.40351
J	-0.07018	0.40351	-0.01754	0.35088	-0.07018	0.40351
A	-0.07018	0.40351	-0.07018	0.40351	-0.07018	0.40351
S	-0.07018	0.40351	-0.07018	0.40351	-0.07018	0.40351

Table 1C: when three fiscal quarters are available (M=3)

	fiscal pattern (a)			fiscal pattern (b)		
t\m	F M A	M J J	A S O	M A M	J J A	S O N
J	0.40676	-0.08889	0.01546	0.40676	-0.08889	0.01546
F	0.40676	-0.08889	0.01546	0.40676	-0.08889	0.01546
M	0.35169	-0.02222	0.00386	0.40676	-0.08889	0.01546
A	0.24155	0.11111	-0.01932	0.35169	-0.02222	0.00386
M	0.07633	0.31111	-0.05411	0.24155	0.11111	-0.01932
J	-0.02222	0.37778	-0.02222	0.07633	0.31111	-0.05411
J	-0.05411	0.31111	0.07633	-0.02222	0.37778	-0.02222
A	-0.01932	0.11111	0.24155	-0.05411	0.31111	0.07633
S	0.00386	-0.02222	0.35169	-0.01932	0.11111	0.24155
O	0.01546	-0.08889	0.40676	0.00386	-0.02222	0.35169
N	0.01546	-0.08889	0.40676	0.01546	-0.08889	0.40676
D	0.01546	-0.08889	0.40676	0.01546	-0.08889	0.40676

Table 1D: when four fiscal quarters or more are available ($M \geq 4$)

	fiscal pattern (a)				fiscal pattern (b)			
t\m	F M A	M J J	A S O	N D J	M A M	J J A	S O N	D J F
J	0.40692	-0.08980	0.01962	-0.00341	0.40692	-0.08980	0.01962	-0.00341
F	0.40692	-0.08980	0.01962	-0.00341	0.40692	-0.08980	0.01962	-0.00341
M	0.35173	-0.02245	0.00491	-0.00085	0.40692	-0.08980	0.01962	-0.00341
A	0.24135	0.11225	-0.02453	0.00427	0.35173	-0.02245	0.00491	-0.00085
M	0.07577	0.31430	-0.06868	0.01194	0.24135	0.11225	-0.02453	0.00427
J	-0.02245	0.37909	-0.02821	0.00491	0.07577	0.31430	-0.06868	0.01194
J	-0.05332	0.30662	0.09689	-0.01685	-0.02245	0.37909	-0.02821	0.00491
A	-0.01685	0.09689	0.30662	-0.05332	-0.05332	0.30662	0.09689	-0.01685
S	0.00491	-0.02821	0.37909	-0.02245	-0.01685	0.09689	0.30662	-0.05332
O	0.01194	-0.06868	0.31430	0.07577	0.00491	-0.02821	0.37909	-0.02245
N	0.00427	-0.02453	0.11225	0.24135	0.01194	-0.06868	0.31430	0.07577
D	-0.00085	0.00491	-0.02245	0.35173	0.00427	-0.02453	0.11225	0.24135
J	-0.00341	0.01962	-0.08980	0.40692	-0.00085	0.00491	-0.02245	0.35173
F	-0.00341	0.01962	-0.08980	0.40692	-0.00341	0.01962	-0.08980	0.40692
M	-0.00341	0.01962	-0.08980	0.40692	-0.00341	0.01962	-0.08980	0.40692

6. EXAMPLES OF THE LOGARITHMIC VARIANT

In order to test the approach to calendarization, the logarithmic variant of section 4 is applied to ten monthly Canadian Retail Trade series. The method is applied in a 5 calendar quarter moving manner as described in section 5. The series, ranging from January 1986 to December 88, were collapsed into fiscal quarter values covering February to April, May to July, etc. The monthly values were then recovered, as interpolations, by applying the method to the fiscal quarter benchmarks, F, and to a seasonal-trading-day pattern, S. For each series, S had been calculated by the X-11-ARIMA seasonal adjustment method (Dagum, 1980), applied to the monthly values. Normally, S would originate from another source.

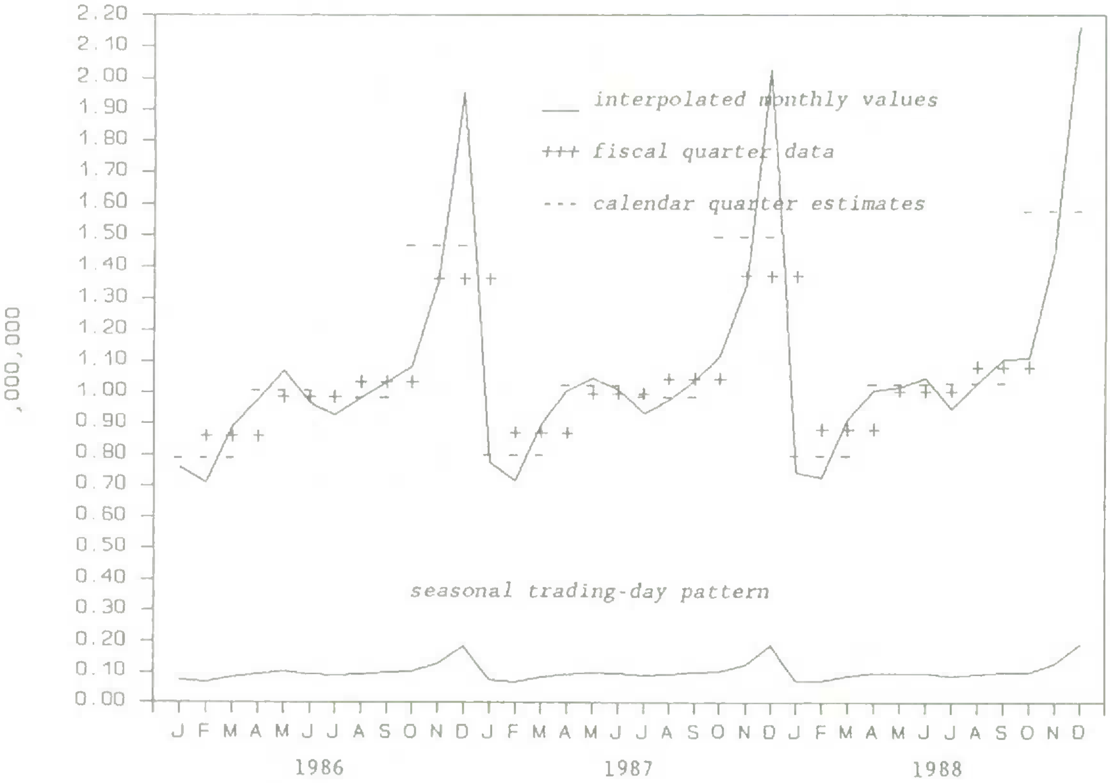


Figure 2: Interpolated monthly estimates and calendar quarter estimates (---) obtained from the logarithmic variant applied to the fiscal quarter data (+++) and to the seasonal-trading-day pattern displayed

The case of Department Stores is illustrated in Figure 2. The interpolations Γ^*_t adopt the month-to-month growth rate of the seasonal pattern and exactly conform to the fiscal quarter benchmarks F_m. This conformity also hold for the calendar quarter estimates, since they are defined as the quarterly sums of the interpolations. Note that most department stores have a fiscal year ending in January and that the fiscal quarter pattern of the figure corresponds to that fiscal year.

Table 2 and 3 present the results for the ten series considered. Table 2A presents statistics on the absolute percentage errors (APE) of the 36 interpolations, with respect to the true monthly values. The low values of the means and standard deviations of the APE's show, in many cases, a surprising degree of accuracy. This demonstrates the possibility of obtaining fairly accurate sub-annual interpolations from a mere seasonal pattern and benchmarks. The table also displays the standard deviations of the seasonal pattern and of the irregular component, σ_s and σ_I, estimated by X-11-ARIMA. Not surprisingly, the accuracy (low statistics) is generally negatively correlated with the intensity of the irregular component (measured by σ_I) of each series.

Table 2B presents the same statistics for the three last interpolations. Contrary to expectations (see Section 5.3), these do not appear ostensibly less accurate than the others. This is due to the fact that, in the fourth quarter of 1988, none of the series displayed a trend-cycle turning-point. However the minimum APE's always differ between Table 2B and 2A, indicating that the precision never reaches its maximum during the last three months. On the contrary, the maximum APE's coincide for 3 of the 10 series, indicating that precision often reach it minimum during the last three months.

Table 2

Analysis of the absolute percent interpolation errors

Table 2A: for the 36 observations of the series

	mean	std. dev.	min.	max.	σ_s	σ_I
Groceries and Meat Stores	0.8	0.7	0.1	2.6	5.9	1.1
Department Stores	0.8	0.7	0.1	2.3	27.6	1.6
General Merchandise Stores	2.6	1.9	0.1	7.2	18.0	3.1
General Stores	1.0	0.7	0.0	3.0	10.8	1.8
Variety Stores	2.1	2.7	0.0	12.5	29.1	2.3
New Vehicles Dealers	2.0	1.5	0.2	6.8	14.1	3.6
Used Car Dealers	2.0	2.2	0.0	10.1	15.8	4.5
Service Stations	0.7	0.6	0.0	2.6	7.0	1.4
Garages	1.1	0.8	0.0	3.2	7.5	2.6
Automotive Parts	1.9	1.5	0.1	5.7	20.5	2.8

Table 2B: for the last three months of the series

	mean	std dev.	min.	max.
Groceries and Meat Stores	1.1	0.3	0.8	1.5
Department Stores	0.5	0.3	0.3	1.0
General Merchandise Stores	2.4	0.9	1.2	3.5
General Stores	1.6	1.1	0.2	3.0
Variety Stores	2.1	1.6	0.9	4.4
New Vehicles Dealers	3.1	2.6	1.2	6.8
Used Car Dealers	1.9	0.9	0.6	3.0
Service Stations	1.4	0.9	0.7	2.6
Garages	1.4	1.1	0.3	2.8
Automotive Parts	3.5	1.5	1.9	5.5

Table 3

Analysis and comparison of the absolute percentage calendarization errors

Table 3A: for the first 11 calendar quarters of the series, under the proposed method and when assigning the fiscal quarters to the closest calendar quarter in brackets

	mean		std dev.		minima		maxima		σ_s	σ_I
Groceries and Meat Stores	0.6	(1.3)	0.4	(0.6)	0.0	(0.4)	1.1	(2.3)	5.9	1.1
Department Stores	0.3	(6.2)	0.2	(3.0)	0.0	(1.6)	0.8	(10.3)	27.6	1.6
General Merchandise Stores	1.1	(7.2)	0.7	(3.4)	0.2	(2.2)	2.6	(12.5)	18.0	3.1
General Stores	0.5	(3.2)	0.3	(1.7)	0.0	(0.4)	1.2	(6.5)	10.8	1.8
Variety Stores	1.5	(5.4)	1.6	(4.6)	0.1	(0.3)	4.6	(12.1)	29.1	2.3
New Vehicles Dealers	0.7	(7.6)	0.3	(6.2)	0.2	(0.1)	1.2	(17.6)	14.1	3.6
Used Car Dealers	0.8	(6.9)	0.8	(6.5)	0.0	(0.4)	2.9	(17.5)	15.8	4.5
Service Stations	0.4	(2.5)	0.2	(1.5)	0.0	(0.5)	0.7	(5.3)	7.0	1.4
Garages	0.6	(3.3)	0.5	(1.9)	0.0	(0.5)	1.5	(6.9)	7.5	2.6
Automotive Parts	1.0	(6.8)	0.8	(5.2)	0.0	(0.3)	2.3	(15.1)	20.5	2.8

Table 3B: APE for the last quarter of each series, under the proposed method

Groceries and Meat Stores	0.6
Department Stores	0.4
General Merchandise Stores	2.4
General Stores	0.6
Variety Stores	1.1
New Vehicles Dealers	1.9
Used Car Dealers	0.7
Service Stations	1.4
Garages	1.3
Automotive Parts	0.2

Table 3A presents the statistics on the APE of the calendarized values, obtained (1) by the method proposed and (2) by assigning, without correction, the fiscal data to the quarter which overlaps the most, in brackets. The means of the APE are from two to twenty times lower with the proposed method. The reduction is especially remarkable for some series with strong seasonality (measured by σ_s), which is not surprising.

Table 3B displays the APE's for the last calendar quarter estimate of the ten series, under the proposed method. The above discussion about Table 2B remains applicable.

The results obtained here are rather encouraging for the proposed method. However, in practice, the seasonal pattern would not be known as precisely. The results presented here may then be interpreted as a sample of the best results that can be expected in real calendarization situations.

7. BACKGROUND

As explained in Section 3, the calendarization method presented in this paper is an adaptation of the benchmarking methods of the Denton type (e.g.: Denton, 1971, Helfand, Monsour and Trager, 1977). The adaptation merely consist of allowing the benchmarks to cover fiscal quarters instead of calendar years. A seasonal-trading-day pattern is then adjusted (benchmarked) to the fiscal quarter benchmarks. The calendarized values are then the calendar quarter sums of the benchmarked series. Cholette and Baldwin (1989) have proposed the same strategy to calendarize fiscal year data; and Cholette and Chhab (1989), to transform aggregates of weekly data into monthly values. The logarithmic variant of Section 4 can be seen as an approximation of the proportional variant of Denton (1971), which is often used in fact as an approximation to a growth rate variant (Smith, 1977).

There is a literature on temporal dis-aggregation. The method by Boot, Feibes and Lisman (1967) coincides with the additive variant of Section 3, if the benchmarks cover calendar years and if the seasonal pattern is quarterly and set to zero. This method is used to convert calendar year data into non-seasonal quarterly values. Cohen, Muller and Padberg (1971) generalized the approach to convert *calendar* data of any frequency into more frequent non-seasonal values.

The temporal disaggregation method proposed by Chow and Lin (1971), Bournay and Laroque (1979), Fernandez (1981), Alba (1988) and others interpolate between benchmarks, by using related series in a linear regression. These methods coincide with additive variant herein, if the benchmarks reflect calendar periods, if only one regressor with coefficient equal to 1 is considered and if the autocorrelation coefficient of the regression residuals is set to 1.

8. CONCLUSION

This paper proposed a method to convert fiscal quarter data into calendar quarter values. The application of the method to a few retail trade series are rather promising. More research would be desirable, especially regarding the preliminary estimates and the use of approximative seasonal patterns.

As explained, the lack of calendarization leads to erroneous data. Yet, to our knowledge, that problem has not drawn the attention of statisticians.

REFERENCES

Alba, E. de (1988), "Temporal Disaggregation of Time Series: a Bayesian Analysis", **Journal of Business and Economic Statistics**, Vol. 6, No. 2, pp 197-206.

Bournay, J., Laroque, G. (1979), «Réflexions sur la méthode d'élaboration des comptes trimestriels», **Annales de l'I.N.S.É.É.**, Vol. 36, pp. 3-30.

Boot, J.C.G., Feibes, W., and Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data", **Applied Statistics**, Vol. 16, no. 1, pp. 65-75

Cholette, P.A. (1988), "Weights to Calendarize Fiscal year Data Referring to Any Consecutive 12 Months or 4 Quarters", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-88-023E.

Cholette, P.A., Baldwin, A. (1989), "Converting Fiscal Year Data into Calendar Year Values", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-89-007E; submitted for publication in **The Journal of Business of Economic Statistics**.

Cholette, P.A., Chhab N. (1988), "Converting Aggregates of Weekly Data into Monthly values", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-89-019E; submitted for publication in **Applied Statistics**.

Cholette, P.A, Dagum, E. Bee (1989), "Benchmarking Socio-Economic Time Series Data: A Unified Approach", **Statistics Canada**, Time Series Research and Analysis Division, Working Paper No. 89-006E.

Chow, G.C., Lin, An-Loh (1971), "Best linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series", **Review of Economics and Statistics**, Vol. 53, No. 4, pp. 372-375.

Cohen, K.J., Müller, W., and Padberg, M.W. (1971) "Autoregressive Approaches to the Disaggregation of Time Series Data", **Applied Statistics**, Vol. 20, pp 119-129.

Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", **Journal of the American Statistical Association**, Vol. 66, No. 333, pp. 99-102.

Dagum, E. (1980), The X-11-ARIMA Seasonal Adjustment Programme, Statistics Canada, Cat. 12-564E; La Méthode de désaisonnalisation X-11-ARIMA, **Statistique Canada**, Cat. 12-564F.

Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series", **Review of Economic and Statistics**, Vol. 63, pp. 471-476.

Helfand, S.D., Monsour, J.J., Trager, M.L., "Historical Revision of Current Business Survey Estimates", **Proceedings of the Business and Economic Statistics Section**, American Statistical Association, 1977.

Smith, P. (1977), "Alternative Methods for Step-Adjustment", **Statistics Canada**, Econometrics Section, Current Economic Analysis Division (internal document).

Young, A.H. (1965), "Estimating Trading-Day Variations in Monthly Economic Time Series", U.S. Bureau of the Census, Technical Paper No. 12.

PART 5

EPIDEMIOLOGY

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

ADJUSTMENT FOR REPORTING-DELAY OF AIDS, AND ESTIMATION OF THE SIZE OF THE HIV INFECTED POPULATION IN THE U.S.A.

I. B. MacNeill[1], Q. P. Duong[2], V. K. Jandhyala[3] and L. Liu[4]

ABSTRACT

The adjustment to the diagnosed AIDS time series to account for reporting-delay is shown to be a function satisfying certain multivariable multiplicative functional equations. Solutions to these equations are characterized for both stationary and non-stationary cases. Estimation of initial conditions is discussed in the context of the U.S. AIDS epidemic. A discussion is given of smoothing and short-term extrapolation of the adjusted series. Following a review of the incubation time distribution for the HIV infection, an integral equation is given which relates the rates for new diagnosed AIDS cases to new HIV infections by means of the incubation time distribution. Solutions of this equation yield estimates of the size of the HIV infected population that are smaller that those previously reported. The most significant feature of the HIV infection estimates is the rapid increase in the rate of infection prior to 1985/86 and the equally precipitous decline after 1985/86; this phenomenon appears to be robust against substantial changes in incubation time distribution and estimates of the size of the diagnosed AIDS population. A discussion is given of implications for the longer term course of the disease.

KEY WORDS: reporting-delay adjustment; short-term extrapolation; integral equations; AIDS forecasting; HIV infection estimation.

1. INTRODUCTION

Reports from the Surgeon General (1986) contained estimates that in 1985 between one and 1.5 million citizens of the USA were HIV infected; these estimates were based on small samples of the general population. Difficulties in obtaining reliable sample survey data in this area make these estimates highly speculative. More recently (1989) the Surgeon General's office has estimated the size of the HIV infected population to be one million; this is a substantial reduction from the earlier estimates, particularly in view of the three year interval between estimates. This paper uses estimates of the rates of progression of the HIV infection to AIDS and data regarding the number of diagnosed cases of AIDS in the USA to obtain alternative estimates of the number with HIV infection. A methodology is developed to account for late-reporting of AIDS data. The alternative estimates of this paper place the size of the HIV infected population at levels lower than that given by the Surgeon General.

2. REPORTING DELAY

Since the AIDS epidemic is now in a crucial phase of its growth curve, it is important for short-term forecasting purposes to have good information regarding the actual number of AIDS cases diagnosed each month. However, after an AIDS case is diagnosed in the U.S. a report of the case must pass through a bureaucratic channel before it reaches the Centers for Disease Control (CDC). The length of time required for a report to reach the CDC is variable. Reports for only a few cases for a particular month will be in the hands of the CDC within the first few months of diagnosis. Most of the cases are reported within the 12 month period following diagnosis. However, some cases may be reported as diagnosed in a particular month as much as several years later; changes in definition of the disease exacerbate the problem.

This late-reporting problem is illustrated by the data graphed in Figure 1. The data are taken from the semi-annual reports issued by the CDC dated January 1987, July 1987, January 1988, July 1988 and January

[1] Department of Statistical and Actuarial Sciences and Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Canada N6A 5B9

[2] Bureau of Management Consulting, 364 Laurier Avenue W. Ottawa, Canada K1A 0S5

[3] Department of Mathematics, Washington State University, Pullman WA 99163

[4] Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Canada N6A 5B9

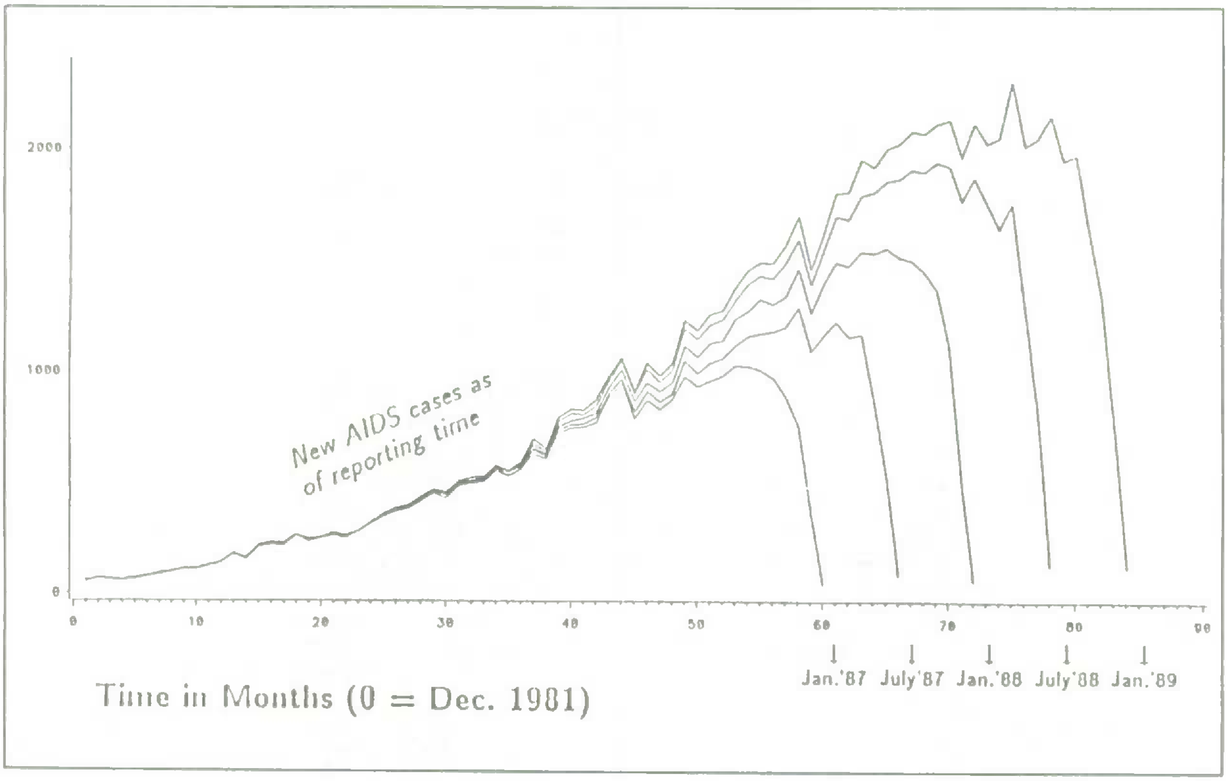


Figure 1. Number of diagnosed AIDS cases reported to the CDC for selected reporting times.

1989, and represent the monthly diagnosed AIDS counts reported to the CDC as of report time. Recent AIDS counts for months in the distant past appear to be approaching close to the totality of diagnosed cases, since all five line-graphs lie close together, but the counts for the more recent past obviously suffer from serious reporting-delay.

The problem addressed in the sequel is that of estimating the total number of AIDS cases diagnosed in a particular month given a partial history as exemplified in Figure 1. Approaches to dealing with the late-reporting problem include those of Morgan and Curran (1986) and Healy and Tillett (1988). We propose a new approach.

Figure 2 is a schematic illustration of the late reporting phenomenon. We let $D_l(t)$ represent the number of new diagnosed AIDS cases for time period t as reported at time l. In Figure 2, $D_l(l-n)$ is the number reported now (l) for n months ago, and $D_{l+m}(l-n)$ is the number that will be reported in m months time for the same month. The reporting-delay adjustment is

$$f_l(n,m) = \frac{D_{l+m}(l-n)}{D_l(l-n)} \quad .$$

We seek $f_l(n,\infty)$ since this represents the reporting-delay adjustment that should be applied to $D_l(l-n)$ to account for all the diagnosed cases that will eventually be reported for the $(l-n)^{\text{th}}$ month. However, we do not wish to wait until m becomes this large.

It is easy to show that $f_l(n,m)$ satisfies the following functional equation:

$$f_l(n, m_1+m_2) = f_l(n,m_1) f_{l+m_1}(n+m_1, m_2) \quad . \tag{1}$$

Iteration of (1) yields

$$f_l(n,m) = \prod_{j=0}^{m-1} f_{l+j}(n+j, 1) \quad .$$

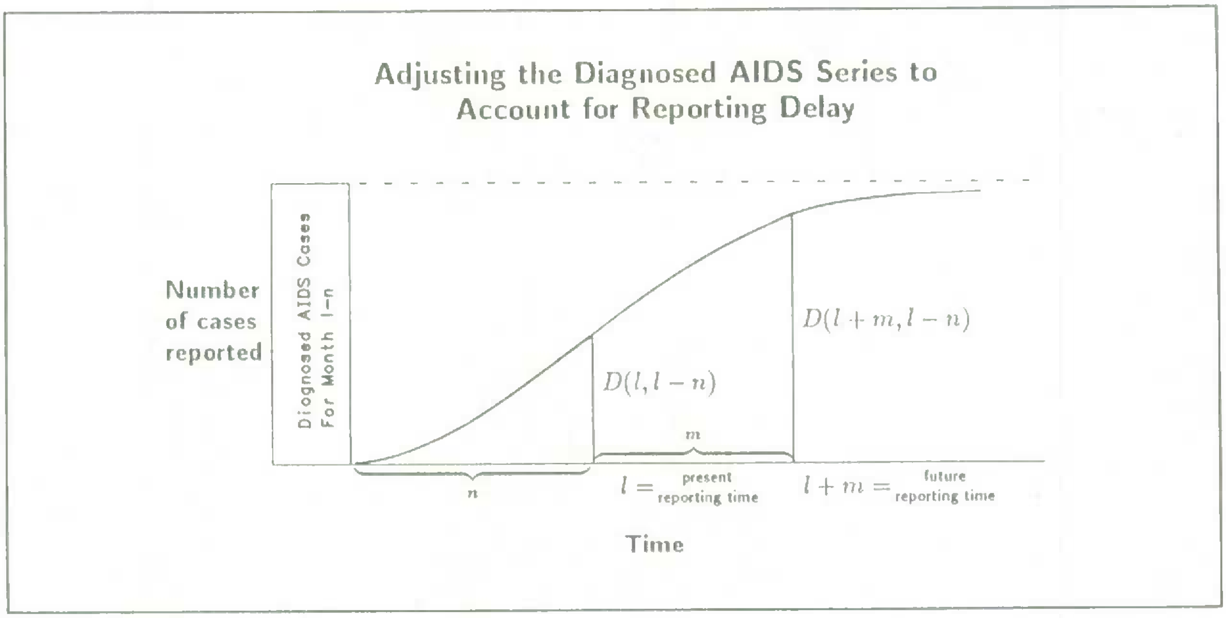

Adjusting the Diagnosed AIDS Series to Account for Reporting Delay
Number of cases reported
Diagnosed AIDS Cases For Month l−n
$D(l+m, l-n)$
$D(l, l-n)$
m
n
$l =$ present reporting time
$l+m =$ future reporting time
Time

Figure 2. Late-reporting for diagnosed AIDS cases.

Hence, if one is given the initial condition $f_l(n,1)$ for all n and l, then one can obtain $f_l(n,m)$ for all l, n, m. This would solve the non-stationary reporting-delay problem since

$$\hat{D}_{l+\infty}(l-n) = f_l(n,\infty) D_l(l-n) \quad .$$

However, estimation of this initial condition places heavy demands on the available data.

If $f_{l_1}(n,m) = f_{l_2}(n,m)$ for all reporting times l_1, l_2, then the functional equation (1) reverts to the stationary case which is stated as follows:

$$f(n, m_1 + m_2) = f(n, m_1) f(n + m_1, m_2) \quad . \tag{2}$$

Again, iteration of (2) yields

$$f(n,m) = \prod_{j=0}^{m-1} f(n+j, 1) \quad .$$

Iteration of (2) also yields

$$f(n, 6m) = \prod_{j=0}^{m-1} f(n+6j, 6) \quad . \tag{3}$$

This initial condition is simpler and, provided that reporting efficiency has remained relatively constant, it may be estimated from the data given in the appendix which is graphed in Figure 1. For the data in Figure 1 (and the appendix), the time unit is the month, with $t = 0$ and $l = 0$ each corresponding to December 1981. For example, $D_{73}(21) = 264$ is the number of diagnosed cases for September 1983 as reported in January 1988. The only reporting dates considered are those from Figure 1, namely $l = 61, 67, 73, 79, 85$. The last months for which data are available are $t = 60$ for $l = 61$, $t = 66$ for $l = 67$, etc. As an example of the adjustment ratio from the data consider $l = 61$, $m = 12$, and $n = 40$. Then $f_{61}(40, 12) = 264/259 = 1.0193$.

The ratio $f_l(m,n)$ is the adjustment to the AIDS count for n months in the past as reported at time l that is required to reproduce the AIDS count for the same month that will be reported m months in the future

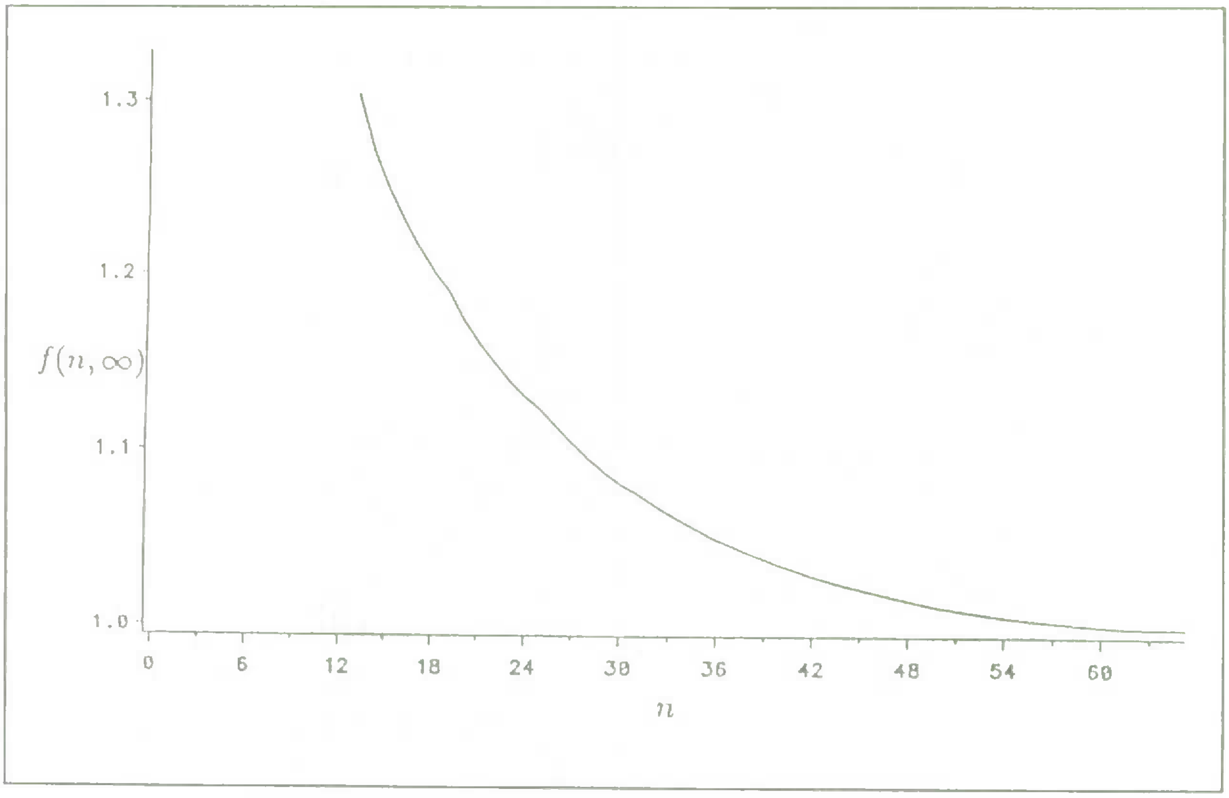


Figure 3. Reporting-delay function, $f(n, \infty)$, $n = 13, 14, \ldots, 60$.

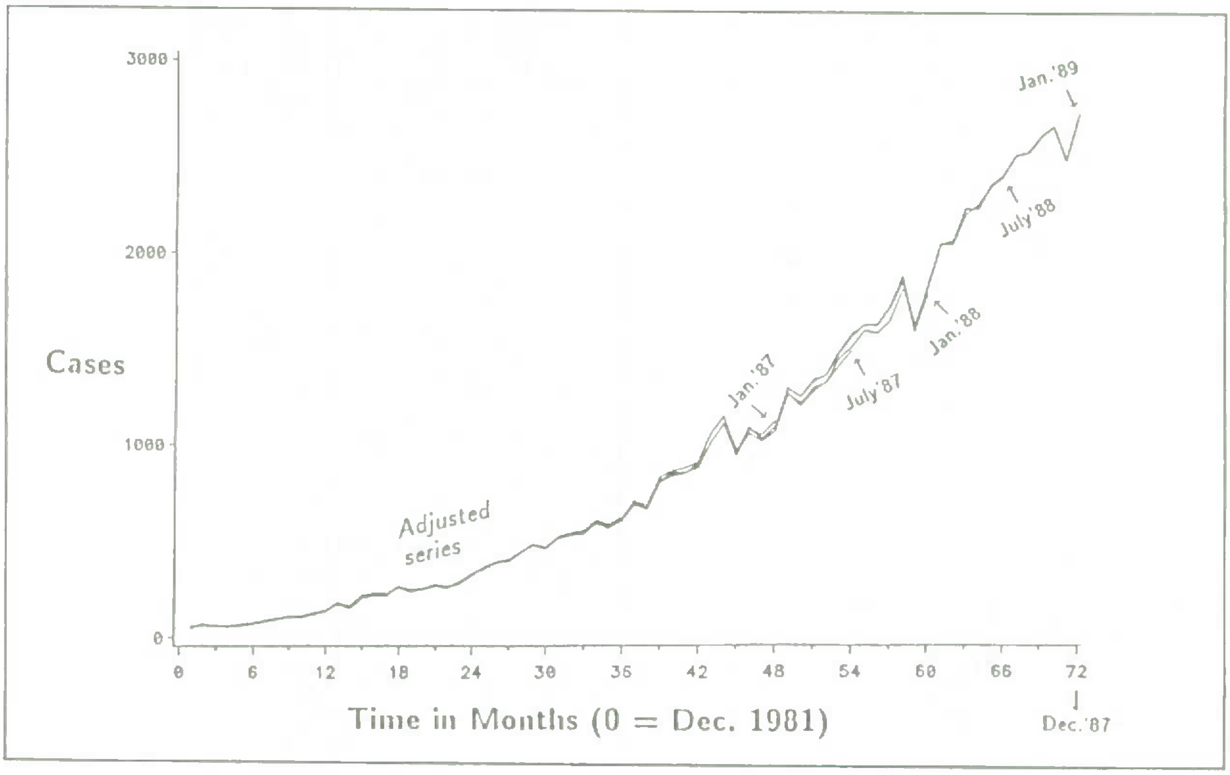


Figure 4. U.S. diagnosed AIDS series as reported January 1989 adjusted for late-reporting, January 1982–December 1987.

(from l). Given a reporting regime with a constant efficiency, then $f_{61}(n,6)$, $f_{67}(n,6)$, $f_{73}(n,6)$, $f_{79}(n,6)$ will each estimate the same ratio. For example, $f_{61}(35,6) = 1.014$, $f_{67}(35,6) = 1.018$, $f_{73}(35,6) = 1.021$ and $f_{79}(35,6) = 1.013$. Morgan and Curran (1986) provide an analysis indicating the reporting delays in the U.S. had not changed significantly up to 1986. Hence we let

$$\hat{f}(n,m) = \operatorname*{ave}_{l}\{f_l(n,m)\} \quad , \tag{4}$$

where l ranges over those report dates for which $f_l(n,m)$ is computable from the available data. For example, using the data in the appendix,

$$\hat{f}(n,6) = \frac{1}{4}\{f_{61}(n,6) + f_{67}(n,6) + f_{73}(n,6) + f_{79}(n,6)\}$$

and more specifically, $\hat{f}(35,6) = 1.0161$. Obviously $f(n,0) \equiv 1.0$.

We note from equation (3) that only the initial condition needs to be estimated from the data. Hence (4) was applied only to $f(n,6)$ for $n = 13,14,\ldots,60$; the data are considered inadequate for $n = 1,2,\ldots,12$. Then the estimates were smoothed and (3) was used to produce $f(n,\infty)$ which is graphed in Figure 3; eight terms in (3) were required to achieve convergence; i.e. $f(n,60) \equiv f(n,\infty)$. Application of this adjustment to the diagnosed AIDS series for January 1982 to December 1987 as reported January 1989 yields the adjusted series as graphed in Figure 4. Application of the adjustment to the diagnosed AIDS series for each of the five reporting times of Figure 1 yields five estimates of the adjusted series. A measure of the validity of the adjustment and of the stationarity of the reporting regime is the degree to which the five adjusted curves coincide. Figure 5, in which the five curves are graphed, indicates the adjustment is working well.

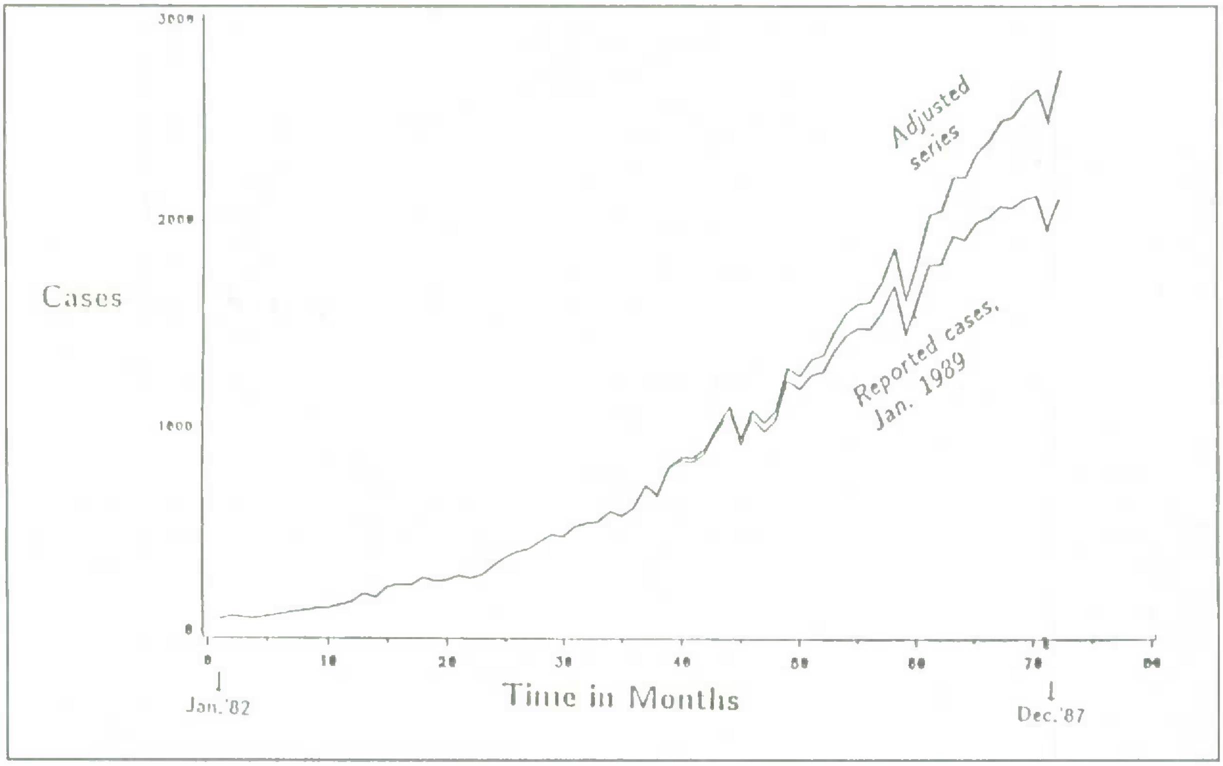


Figure 5. Adjusted U.S. diagnosed AIDS series for five selected reporting times.

By way of comment it can be noted that (1) and (2) are discrete cases of functional equations that generalize the well known functional equation

$$f(t_1 + t_2) = f(t_1)f(t_2) \quad ,$$

whose solution under mild regularity condition is the exponential. The continuous version of (2) is the bivariate functional equation,

$$f(s,\, t_1 + t_2) = f(s, t_1) f(s + t_1, t_2) \quad .$$

Under mild regularity conditions MacNeill (1989) has characterized the non-trivial solutions as

$$f(s,t) = \exp\left\{\int_0^t g(s+x)\,dx\right\}$$

where the initial condition is

$$g(s) = \frac{\partial}{\partial t} f(s,t)\Big]_{t=0} \quad .$$

The equation $f(s,t)$ is called the stationary reporting-delay function. The non-stationary reporting-delay function satisfies the following equation:

$$f_l(s, t_1 + t_2) = f_l(s, t_1) f_{l+t_1}(s + t_1, t_2) \quad .$$

Again, under mild conditions, the non-trivial solutions are characterized as follows:

$$f_l(s,t) = \exp\left\{\int_0^t g(l+x, s+x)\,dx\right\}$$

where the initial condition is

$$g(l,s) = \frac{\partial}{\partial t} f_l(s,t)\Big]_{t=0} \quad .$$

Other functional equations exhibiting this multiplicative property are discussed by MacNeill (1989).

3. RELATIONSHIP BETWEEN DIAGNOSED CASES, INFECTIONS AND INCUBATION TIME

Our purpose in this section is to develop equations relating the number of AIDS cases diagnosed in a certain jurisdiction per time unit to the number who acquired the HIV infection in prior time units. For the sake of argument we use years as time units. We let $D(k)$ represent the number diagnosed with AIDS in year k. Also, we let $I(j,k)$ represent the number infected in year j and diagnosed with AIDS in year $k (k \geq j)$; then

$$D(k) = \sum_{j='76}^{k} I(j,k) \quad .$$

We assume here, for the sake of argument, that the first infections occurred in 1976 or later. Thus, if $P(j,k)$ represents the proportion of the total infected in year j that is subsequently diagnosed with AIDS in year k, and if $N(j)$ is the total number infected in year j, then

$$P(j,k) = \frac{I(j,k)}{N(j)} \quad .$$

We let $T(l)$ be the number infected to year l; these are the quantities about which so little seems to be known and for which we can provide estimates using the following system of equations:

$$
\begin{aligned}
T(l) &= \sum_{j='76}^{k} N(k) \qquad l =' 76,' 77, \ldots \quad , \\
D(k) &= \sum_{j='76}^{k} N(j)\, P(j,k) \qquad k =' 76,' 77, \ldots \quad .
\end{aligned}
\tag{5}
$$

As discussed in Section 2, substantial information is now available regarding $D(k)$, and other estimates can be obtained by forecasting; this time series is discussed further in Section 4. One of the first studies to provide information regarding $P(j,k)$ comes from the clinical work of Brodt et al (1986), and from the analysis of their data by Cowell and Hoskins (1987) and Panjer (1987). More recent studies reported by Kalbfleisch and Lawless (1988) yield other estimates of incubation times. Section 5 contains a discussion of the estimation of $P(j,k)$. In Section 6 the methodology of this section is applied to the series for diagnosed cases and to the incubation time distribution to obtain estimates of the size of the HIV infected population in the USA.

Meanwhile, it can be noted that equation (5) is the discrete analogue of the following integral equation,

$$
D(t) = \int_0^t P(t-s) N(s)\, ds \quad . \tag{6}
$$

Hence, if $D(\cdot)$ and $P(\cdot)$ are known, one may obtain $N(\cdot)$ by solving (6). This provides a powerful tool for studying the relationships among plausible models for the diagnosed rate function $D(\cdot)$, the infection rate function $N(\cdot)$ and the incubation time distribution $P(\cdot)$.

If $\mathcal{L}(f)$ represents the Laplace transform of the function $f(\cdot)$, then equation (6) yields

$$
\mathcal{L}(D) = \mathcal{L}(P)\mathcal{L}(N) \quad .
$$

If $D(\cdot)$ and $P(\cdot)$ are known and their transforms can be obtained analytically then

$$
\mathcal{L}(N) = \mathcal{L}(D)/\mathcal{L}(P) \quad .
$$

The infection distribution $N(\cdot)$ may then be obtained by inverting its transform. As an example, consider the functions

$$
D_1(t) = K \exp\{-c/t\} \ , \qquad t > 0 \ ,
$$

and

$$
P_1(t) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha}\, t^{\alpha-1} \exp\{-t/\beta\} \ , \qquad t > 0 \quad .
$$

This choice of diagnosed rate function implies growth from zero at time $t = 0$ to the asymptote K with the rate of growth controlled by c. Qualitatively, such growth is epidemiologically plausible for the time frame in which we are presently interested.

As a concrete case with gamma distributed incubation times consider $\alpha = 2$ and $\beta = 4$, which implies a mean incubation time of 8 years. The Laplace transforms of $D_1(t)$ and $P_1(t)$ are:

$$
\mathcal{L}(D_1) = 2K\sqrt{c}\, p^{-\frac{1}{2}} K_1(2\sqrt{c}\, p^{\frac{1}{2}}) \ ,
$$

where $K_1(\cdot)$ is a modified Bessel function, and

$$
\mathcal{L}(P_1) = \frac{1}{16}(p + \frac{1}{4})^{-2} \quad .
$$

Then,

$$
\mathcal{L}(N_1) = 32K\sqrt{c}\, p^{-\frac{1}{2}} K_1(2\sqrt{c}\, p^{\frac{1}{2}})(p + \frac{1}{4})^2 \quad ,
$$

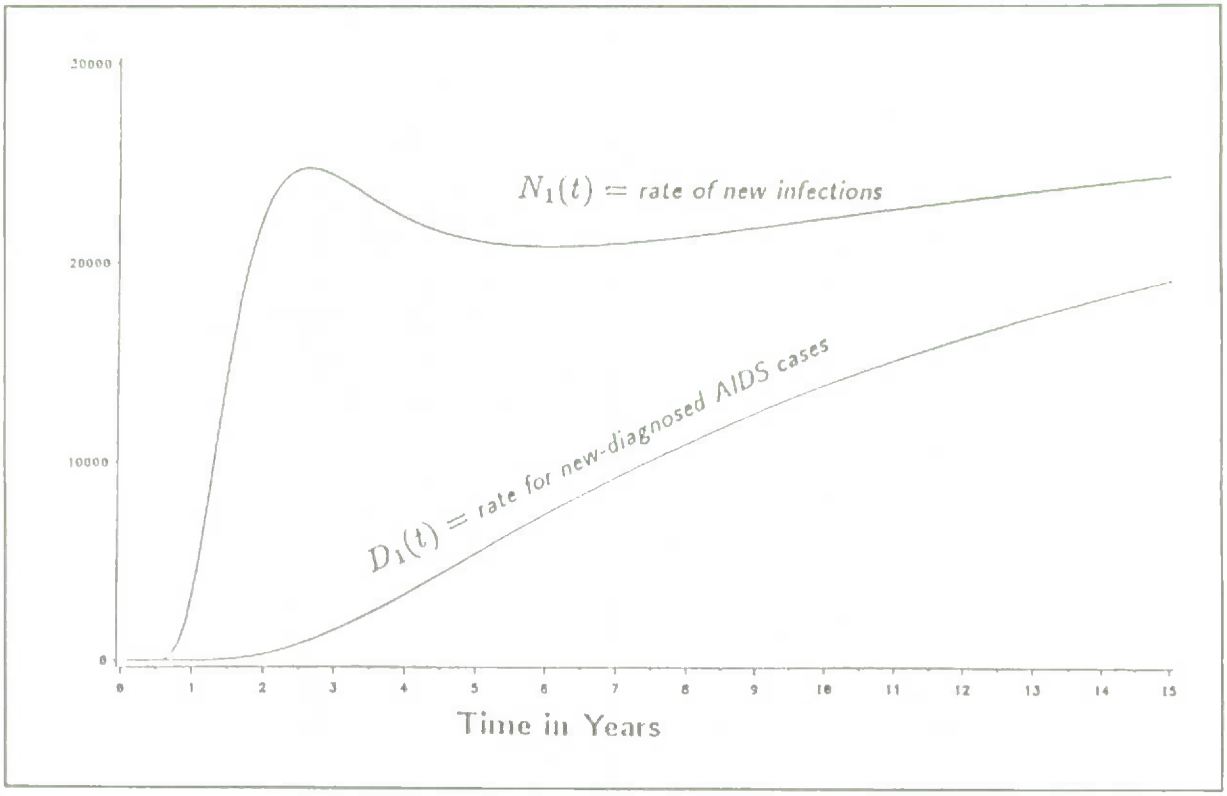


Figure 6. Rates for new diagnosed AIDS cases and new HIV infections; model 1.

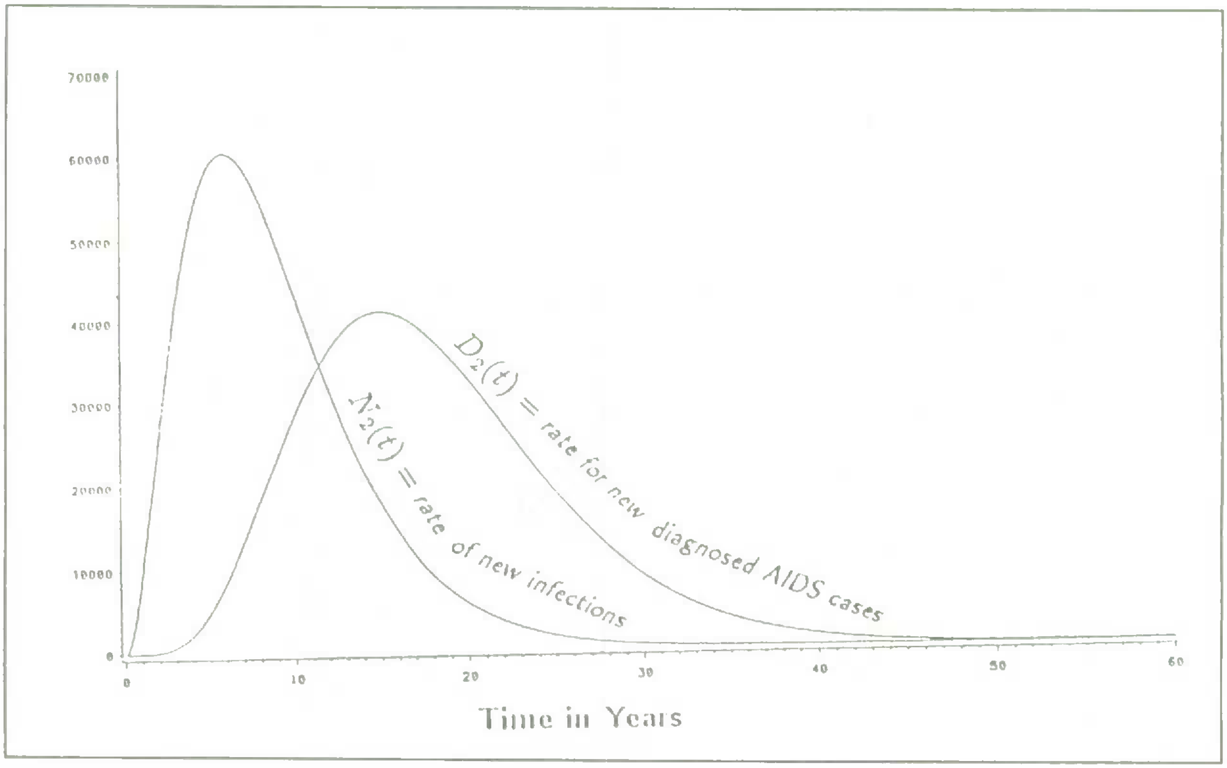


Figure 7. Rates for diagnosed AIDS cases and new HIV infections: model 2.

inversion of which yields

$$N_1(t) = K \exp\{-c/t\}(1 + 8ct^{-2} - 32ct^{-3} + 16c^2t^{-4}) \quad .$$

Figure 6 presents $D_1(t)$ and $N_1(t)$ with $K = 36,000$ and $c = 9.5$.

As a second example, consider

$$D_2(t) = \frac{K}{\Gamma(\alpha)\ \beta^{\alpha}}\ t^{\alpha-1} \exp\{-t/\beta\} \qquad t > 0$$

and

$$P_2(t) = \frac{1}{\Gamma(\alpha)\ \beta^{a}}\ t^{a-1} \exp\{-t/\beta\} \quad , \qquad t > 0 \quad .$$

This choice of diagnosed rate function implies growth from zero at $t = 0$ to a maximum and then a decline asymptotic with zero; the rates of growth and decline are determined by α and β. Again such growth is qualitatively plausible for certain epidemiological series although it might be premature to forecast the time of the downturn in the number of diagnosed AIDS cases; the question is discussed below in Section 7.

Solution by the Laplace transform method yields

$$N_2(t) = \frac{K}{\Gamma(\alpha - a)\beta^{\alpha - a}}\quad t^{\alpha - a - 1} \exp\{-t/\beta\} \quad .$$

Figure 7 presents $D_2(t)$ and $N_2(t)$ with $\alpha = 6$, $a = 3$, $\beta = 3$ and $K = 699,840$. The mean incubation time implied by this choice of parameters is 9 years.

The parameters for each of the above examples were chosen to yield a total of approximately 80,000 AIDS cases in the first 10 years of the epidemic.

In the event that analytical solutions to the integral equation (6) are not available, one may resort to numerical techniques to obtain solutions. These methods, which are based on the equations in (5), can be tested on the exact solutions represented in Figures 6 and 7. This has been done, and agreement to pre-determined levels of accuracy can be attained.

4. THE NUMBER OF DIAGNOSED CASES OF AIDS

The first diagnosed cases of AIDS in the USA occurred in 1978; earlier cases may have occurred but were undiagnosed. The size of the epidemic increased exponentially for several years thereafter. This early exponential growth provoked forecasts of a calamity that would rival the Black Death, a catastrophe which decimated the population of Europe during the 1300's. The exponential forecasts were made with the proviso that trends present at that time would continue.

However, the growth of the number of diagnosed incidences of AIDS appears to have departed from the exponential mode in early 1984. Duong and MacNeill (1987) identified this departure for the Canadian data, and Jandhyala and MacNeill (1988) have analyzed the U.S. data and have estimated early 1984 as the date at which the parameters of the system begun to change. The methodology used to test the hypothesis of parameter change at unknown time is given by Chernoff and Zacks (1964), MacNeill (1978) and Jandhyala and MacNeill (1986). Having determined that the exponential hypothesis is no longer tenable, Duong and MacNeill (1986) use the Akaike information criterion (AIC) to select from among a range of growth models that which best fits the Canadian data. The AIC can be used to compare non-nested models, and brings model selection into the inferential process. The model chosen for the Canadian data is the logistic, and it can be used to forecast the incidence of AIDS; previous applications of this methodology suggests that, at least for the near future, these forecasts can be expected to be reasonably accurate. At this stage in the development of the epidemic, it is unlikely that realistic non-empirical models with the large number of parameters that they entail will be of much use in near-term forecasting. Hence, we will rely upon the logistic for smoothing and for near-term forecasts of the number of diagnosed AIDS cases.

The logistic function is defined as follows:

$$D(t) = \frac{M D(0)}{D(0) + (M - D(0)) \exp\{-mkt\}} \quad ,$$

where $D(0)$ is the size of the epidemic at $t = 0$, M is the maximum size (rate) of the epidemic, k is the slope factor, and $D(t)$ is the rate of diagnosis at time t. This function has been fitted by non-linear least squares to the monthly diagnosed AIDS cases as reported January, 1989 by the Centers for Disease Control (CDC); December 1981 is regarded as time $t = 0$. The data for 1988 are not used for fitting due to severe late-reporting which characterizes the reporting of the AIDS epidemic. The data prior to 1988 have been revised upwards using the late-reporting adjustment discussed in Section 2. The logistic is used here only for very short-term forecasting, viz, 12 months to December 1988.

Figure 8 shows the graph of the adjusted data superimposed on the fitted logistic curve.

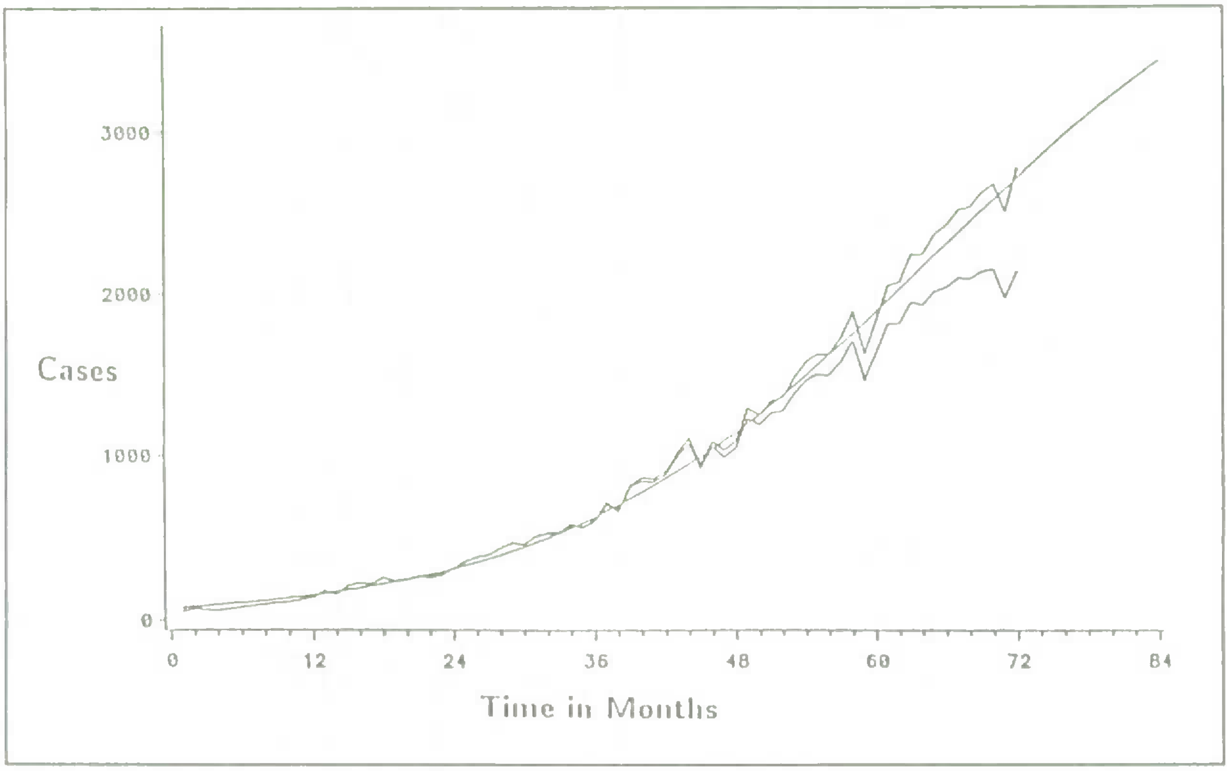


Figure 8. U.S. diagnosed AIDS series (adjusted), January 1982 to December 1987 as reported January 1989, with a logistic fit extrapolated to December 1988.

5. INCUBATION TIMES FOR HIV INFECTIONS

A longitudinal study by Brodt et al. (1986) conducted at the University of Frankfurt followed subjects at risk of AIDS to determine the times of progression through the various stages of the disease. The study used the five stages of the Walter Reed Staging Method to identify progress from healthy status to AIDS. The stages are:

1a (At-Risk): Healthy persons at risk for HIV infection, but testing negative.

1b HIV^+: Otherwise asymptomatic persons testing HIV^+.

2a (LAS): Patients with HIV infection and lymphadenopathy syndrome (LAS), together with moderate cellular immune deficiency.

2b (ARC): Patients with HIV infection and LAS, together with severe cellular immune deficiency (AIDS-related complex, ARC, as defined by CDC).

3 AIDS: Patients with AIDS as defined by CDC.

The last stage is death.

Table 1, which presents some of the main results of the study, gives the number of patients observed by stage and by length of observation period.

Range of Observation Periods	Stage 1a (At-Risk)	Stage 1b (HIV^+)	Stage 2a (LAS)	Stage 2b (ARC)	Stage 3 (AIDS)	All Stages
3-6 months	10	9	21	8	6	54
6-12 months	14	18	51	29	9	121
12-24 months	21	20	29	20	7	97
24-36 months	3	5	19	7	$1^{(1)}$	35
All Periods	48	52	120	64	23	307

Table 1. Frankfurt Study "Table 5" Data: Number of Patients Observed by Stage and Observation Period

n	1	2	3	4	5	6	7	8	9	10	11	12
$P(j, j+n)$	0.016	0.065	0.107	0.125	0.124	0.113	0.096	0.080	0.063	0.050	0.039	0.030

Table 2. Proportion of those newly infected which will have, or will die of, AIDS n years later

These data have been used by Cowell and Hoskins (1987) and by Panjer (1987) to estimate HIV progression rates and AIDS mortality rates. The model developed by Panjer was used to derive the estimates of $P(j, j, +n)$ presented in Table 2. This model implies an incubation time distribution that is approximately gamma with a mean of 6.3 years. Cowell and Hoskins in an analysis of the same data but with a different model obtain a mean that is approximately two years longer. Kalbfleisch and Lawless (1988) have estimated the median incubation time to be approximately 10 years. In the sequel we have used the gamma distribution with various values of the parameters as models for incubation time distribution. We comment later on upper bounds for mean incubation times, and suggest that 10 years is approaching this upper bound.

6. ESTIMATES OF THE SIZE OF THE HIV INFECTED POPULATION IN THE USA

Figure 8 contains the U.S. monthly data regarding diagnosed AIDS cases adjusted for late-reporting. The logistic model fitted to the data and extrapolated forward to the end of 1988 is used in estimating the size of the HIV infected population. These estimates are made using as incubation time distribution the Gamma distribution with the various parameter sets suggested by studies discussed in Section 5. Equation (6) is solved numerically and yields the estimates given in Table 3. Figures 9 contains the graph of $D(t)$, the rate for new diagnosed AIDS (logistic fit), and $N(t)$, the rate of new infections. For this characteristic case, the parameters for the logistic fit are $D(0) = 75.92$, $M = 4476.86 \pm 255.33$ and $k = 0.00001412$. The incubation time distribution assumed for Figure 9 is gamma with $\alpha = 2$, $\beta = 5$ for a mean incubation time of 10 years. The computations are repeated with the asymptote of the logistic curve increased by two standard errors.

The most striking feature of the graph of $N(t)$, the infection rate function, is the rapid rise in the number of new infections until 1985/86 and the equally precipitous decline during subsequent years. As noted these computations have been repeated for a variety of plausible incubation time distributions and for larger asymptotes for the logistic model. In all cases the infection function rises rapidly prior to 1985/86 and declines sharply after 1986. Other factors held constant, longer average incubation times result in larger estimates of the size of the HIV infected population; this is illustrated in Table 3.

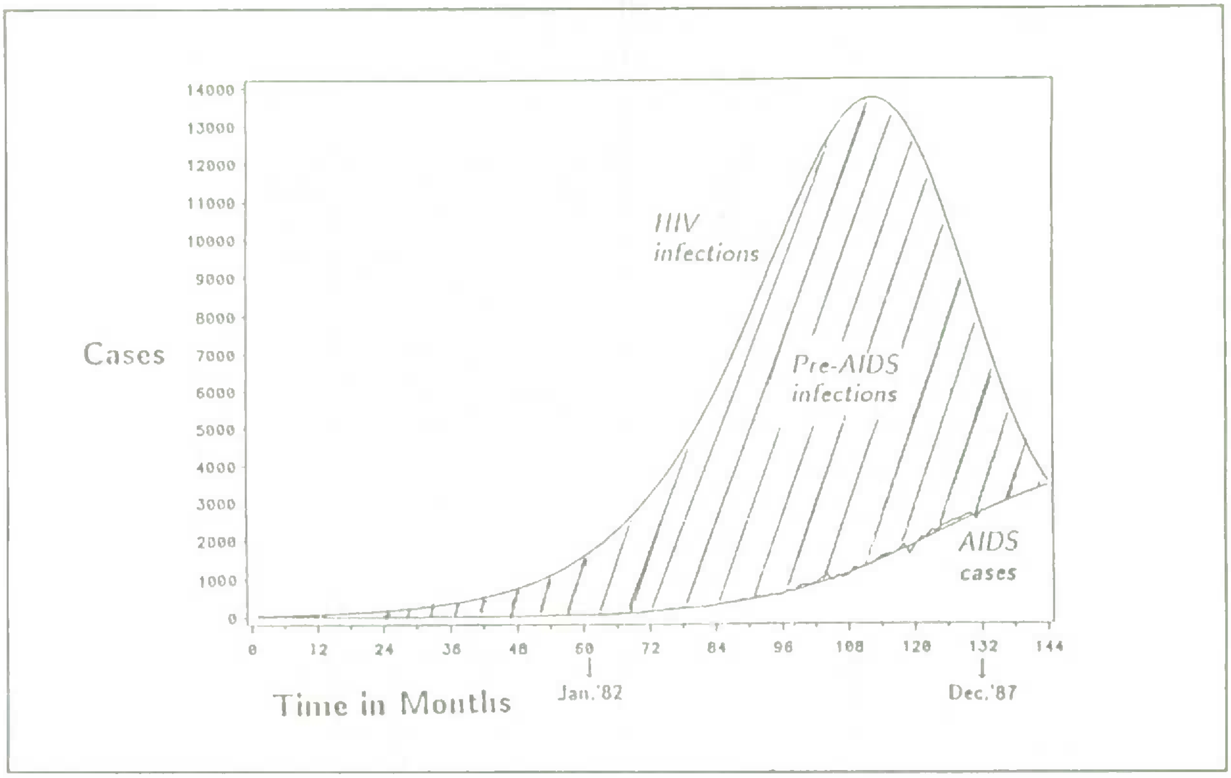


Figure 9. Rate of HIV infection recruitment, $N(t)$, and rate of diagnosis of AIDS, $D(t)$.

Mean Incubation Times	Pre-AIDS HIV infections			
	1985		1988	
	Logistic	Logistic $+2\sigma$	Logistic	Logistic $+2\sigma$
8 years	232,500	466,800	443,300	498,400
9 years	278,800	558,600	515,100	567,500
10 years	329,000	657,700	590,600	638,000

Table 3. Estimates of the size of the pre-AIDS population in the USA

The above analysis suggests that the number of latent AIDS cases in the USA approximates to 600,000, which is a large number but considerably smaller than that estimated by the Surgeon General (1989).

It can be noted that positivity requirements on the $N(t)$ series impose upper bounds on the mean length of the incubation time for a particular model of the distribution. In the above formulation, for incubation time distributions to be consistent with the apparent trend in diagnosed AIDS cases, equation (6) and positivity of $N(t)$ require that the mean incubation time be bounded at not much more than 10 years.

7. AIDS FORECASTS

What does the future hold regarding the number of AIDS cases in the USA? Several scenarios may be explored by extrapolation of $N(t)$, the HIV infection rate function; Figure 10 contains three different extrapolations corresponding to a mean incubation time of 10 years. Equation (6) may then be used to forecast the AIDS series. Extrapolation A is compatible with continued logistic growth in $D(t)$, the diagnosed AIDS case series. Extrapolations B and C are more in keeping with the apparent internal dynamics of the $N(t)$ series, and suggest a decline in the number of diagnosed AIDS cases beginning in 1990. The corresponding forecasts for $D(t)$ based on the three scenarios for $N(t)$ appear in Figure 11.

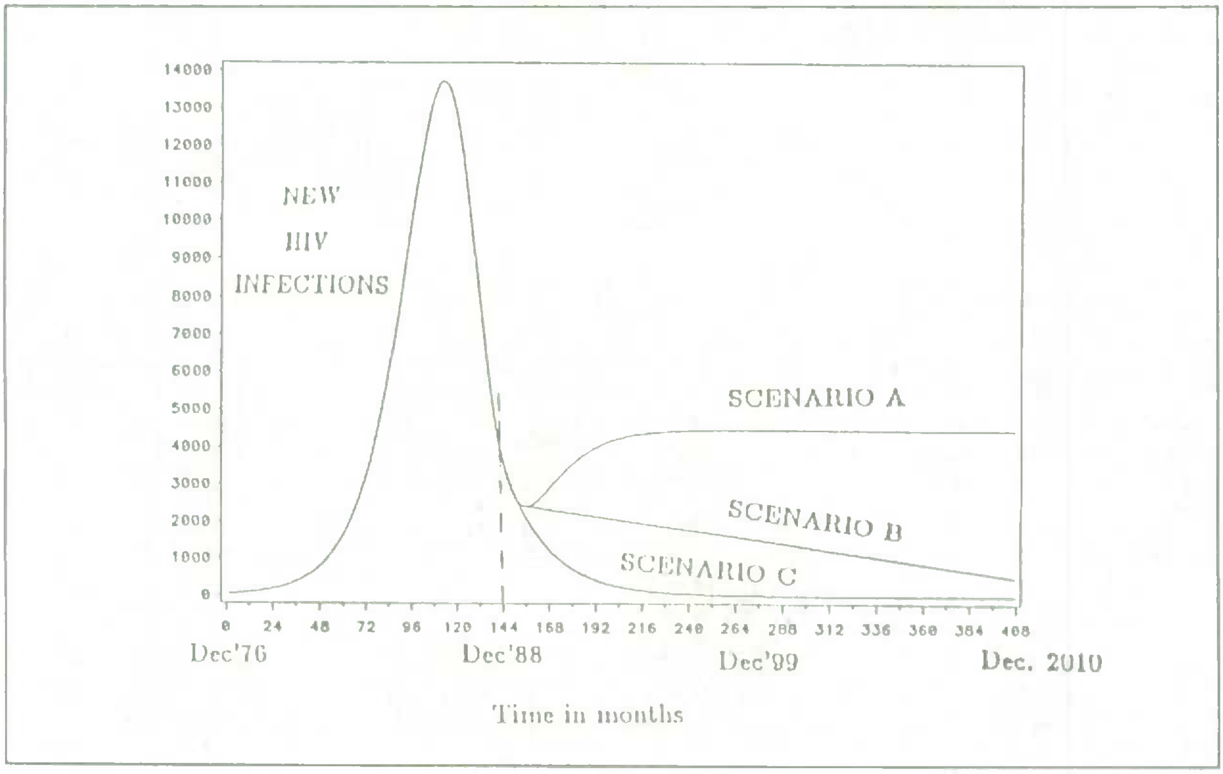


Figure 10. New HIV infections with three scenarios for the future (A, B, C).

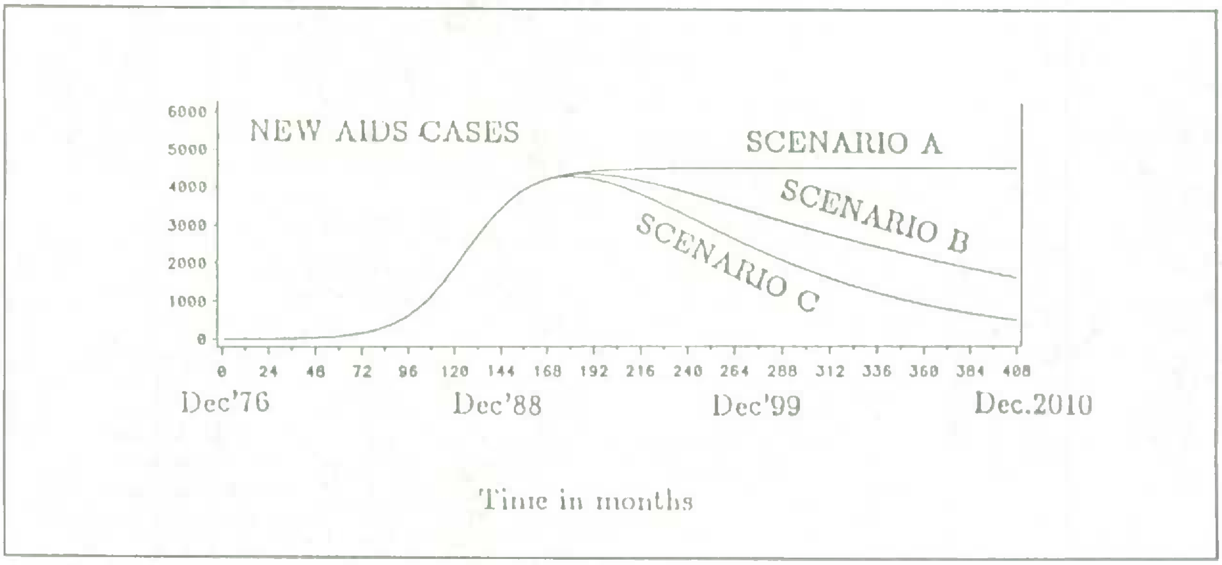


Figure 11. New AIDS cases with three scenarios for the future (also see Figure 10).

8. DISCUSSION

The analysis given above provides estimates of the size of the HIV infected population that are lower than those given in the Coolfont report (1986) and in the more recent reports from the Surgeon General's office.

Several factors significantly affect the size of the estimates given in this paper. First, the longer the incubation time, the larger is the estimate of the size of the HIV infected population. This follows because the number of cases of AIDS observed up to the present represents a proportion of the infected population that varies inversely with the length of time it has taken to develop these AIDS cases.

Second, the percentage of the infected population that ultimately become AIDS victims is assumed to be 100%. If this assumption is not true, and the percentage is 100 p %, where $0 < p < 1$, the estimates of the size of the HIV infected population must be increased by the factor p^{-1}. No adjustment is necessary if one is interested only in latent AIDS cases.

Third, information regarding under-reporting is largely speculative. However, if the fraction of all cases that are reported is f, then the estimates of the size of the HIV infected population must be increased by the factor f^{-1}. One could speculate that the effect of under-reporting is unlikely to exceed two standard errors on the logistic asymptote.

The most significant feature of the HIV infected estimates is that they have peaked and have declined rapidly since 1986. Several reasons may be put forth to explain the post 1986 precipitous decline. First, the effect of education among the high risk groups has resulted in less risky behaviour on their part. However, because of the lengthy mean incubation time, education probably had a minimal impact on the shape of the $N(t)$ function in the pre 1985/86 period. The second, and more plausible, explanation for the rapid growth and decline, is a saturation effect among those who put themselves most at risk during the period immediately prior to the time when general awareness of the disease emerged. That is, exponential-type growth occurred at first but after the infection spread to a large part of this group there was little room left for continued growth.

The education program and the broad dissemination of information about AIDS that occurred during the earlier part of this decade are likely to have their greatest impact on the HIV infection rate during the next decade. If these programs have been effective then scenario C will be more likely; if the programs have not been effective then scenario A will be more likely.

Evidence given by Johnston (1988), McKusick et al (1985), Martin (1987), Winkelstein et al (1987) and others, to the effect that among the homosexual male population in the USA, education has resulted in substantial behavioural change. Hence, the weight of this evidence points to scenario C as the most likely. The effect of education upon IV drug users is more problematical.

However, even if new infections were to cease immediately, AIDS will continue to be a significant epidemic through the 1990's simply because the number of latent cases remains relatively large.

The more encouraging news is that effects of the rapid growth and precipitous decline in the rate of increase of the HIV infected population in the USA prior to the late 1980's will play themselves out during the 1990's and will have a much diminished impact by the turn of the century.

The above discussion assumes no imminent major medical discovery in the form of cures or vaccines for AIDS. It also assumes no discovery of drugs or therapies that would lengthen the mean incubation time of HIV infection and/or the survival time of AIDS victims. Obviously such cures, vaccines, drugs and therapies would have a major impact on the course of the epidemic.

The previous analysis has been predicated upon the notion that the incubation time distribution is stationary. What would be the effect of a lengthening of the mean incubation time, perhaps through drug therapy, at some future point? Figure 12 contains graphs of $D(t)$ for scenarios A, B and C when the mean incubation time is extended from 10 years to 15 years beginning in 1990. The principal effects of a lengthening of the incubation times are to reduce the impact of the epidemic in the 1990's but to increase it in the early part of the next century.

It should be noted that the time from the instant of HIV infection to death from AIDS is divided into two periods by the instant of AIDS diagnosis. The first of these periods is the incubation time, and the second is the survival time. If AIDS diagnoses are postponed, for whatever reason, then incubation times will be artificially lengthened, and survival times will be correspondingly artificially shortened. The opposite would be the case if diagnoses are made earlier than called for by the staging method. If the former had been the case, perhaps due to a desire to avoid negative social effects, and if the latter were now the case, perhaps due to the promise of improved survival from an untested AIDS drug therapy, then spurious and offsetting changes could be noted in incubation and survival times. In fact, it is possible for actual survival times to be shortened by an untested, perhaps expensive, drug therapy and yet appear to be lengthened due to a larger spurious effect of the kind noted above. The regulation of such snake-oil therapies is another argument for conducting properly randomized clinical trials.

9. IN CONCLUSION

If medical science is unsuccessful in discovering cures, vaccines or effective drugs/therapies, then mankind will have to manage with AIDS as it has managed with other epidemics in the past; namely by developing natural immunity. The childhood diseases brought by the Spanish to the New World in the 16-th century had

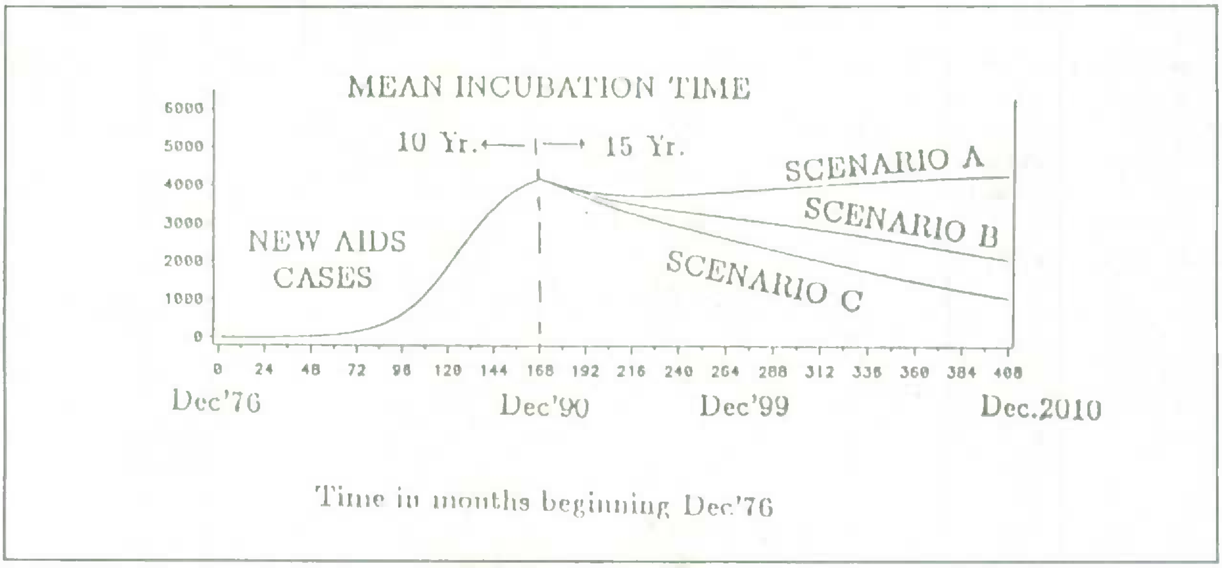


Figure 12. New AIDS cases with three scenarios for the future (see Figure 10) and with an increase in mean incubation time from 10 to 15 years in 1990.

devastating effects upon the native populations of the Americas. However, today these diseases have no more impact upon their descendants than they do upon the descendants of the Spanish. Several generations have been required to build immunity defences in past; W.H. McNeill (1975) has estimated six generations. To the extent that pediatric cases of the HIV infections are relatively rare, it may require more generations than usual to build immunity to AIDS. Meanwhile, education is the main hope for near-term management of the AIDS epidemic.

References

Brodt, H. R., E. B. Helm, A. Joetten, L. Bergmann, A. Kluver, and W. Stille (1986). *Spontanverlauf de LAV HTLV-III-Infektion; Verlaufsbeobachtungen bei Personen aus AIDS-Risikogruppen;* Deutsche Medizinische Wochenscrift, Stuttgart, Vol. iii, pp. 1175-1180.

Chernoff, H., and S. Zacks (1964), "Estimating the current mean of a normal distribution which is subject to changes in time". *Annals of Mathematical Statistics* **35**, 999-1018.

Cowell, M. J., and W. H. Hoskins (1987). AIDS, HIV Mortality and Life Insurance, Parts 1 and 2, Society of Actuaries, distributed as a Special Report.

Duong, Q. P., and I. B. MacNeill (1987). Selection and estimation of growth models with application to forecasting AIDS. *Department of Statistical and Actuarial Sciences, Technical Report TR-87-09.* London, Canada: The University of Western Ontario.

Healy, M. J. R. and H. E. Tillett (1988). Short-term extrapolation of the AIDS epidemic. *Journal of the Royal Statistical Society, Series A*, 50-61.

Jandhyala, V. K. and I. B. MacNeill (1989). Detection of parameter changes at unknown times in linear regression models. *(to appear).*

Jandhyala, V. K., and I. B. MacNeill (1989). Change detection methodology for modelling the incidence of AIDS. *Department of Statistical and Actuarial Sciences, Technical Report TR-89-01.* London, Canada: The University of Western Ontario.

Johnson, A. M. (1988). Social and behavioural aspects of the HIV epidemic—A review. *Journal of the Royal Statistical Society, Series A*, **151**, 99-114.

Kalbfleisch, J. D. and J. F. Lawless (1988). Inference based on retrospective ascertainment. An analysis of the data on transfusion related AIDS. *Technical Report STAT-88-02,* Department of Statistics and Actuarial Science, University of Waterloo.

MacNeill, I. B. (1978). Properties of sequences of partial sums of polynomial regression residuals with application to tests for change in regression at unknown times. *Annals of Statistics* **6**, 422-433.

MacNeill, I. B. (1989). The reporting-delay function. *Department of Statistical and Actuarial Sciences Technical Report TR-89-04.* London, Canada: The University of Western Ontario.

Martin, J. L. (1987) The impact of AIDS on gay male sexual behaviour patterns in New York City. *American Journal of Public Health*, **77**, 578-581.

McKusick, M., W. Horstman, and J. J. Coates (1985). AIDS and sexual behaviour reported by gay men in San Franscisco. *American Journal of Public Health*, **75**, 493-496.

McNeill, W. H. (1976). *Plagues and Peoples.* Doubleday: New York.

Morgan, W. Meade and James W. Curran (1986). Acquired Immunodeficiency Syndrome: current and future trends. *Public Health Report 101* 5, 459-465.

Panjer, H. H. (1987). AIDS: Survival analysis of persons testing HIV$^+$. Working paper series in Actuarial Science ACTSC87-14, Waterloo, Canada: The University of Waterloo.

"Public Health Service Plan for the Prevention and Control of AIDS and the AIDS virus", Report of the Coolfont Planning Conference, U.S. Public Health Service, Washington, 1986, p.1.

Winkelstein, W., M. Samuel, N. S. Padian, and J. A. Whiley (1987). Select sexual practices of San Franscisco heterosexual men and risk of infection by the human immunodeficiency virus. *Journal of American Medical Association*, 257, 1470-1471.

APPENDIX

Month		Date of CDC Report				
		Jan. 87	July 87	Jan. 88	July 88	Jan. 89
Prior to'82		337	341	352	348	381
(Jan.'82)	1	50	50	48	49	54
	2	64	66	68	67	68
	3	58	59	59	59	62
	4	59	58	57	56	58
	5	59	61	61	64	68
	6	71	71	72	74	76
	7	81	84	84	85	88
	8	92	94	95	94	96
	9	109	105	106	107	108
	10	104	106	106	106	111
	11	118	123	121	125	125
	12	134	135	137	137	139
	13	170	174	172	174	180
	14	149	153	151	153	161
	15	200	203	207	212	216
	16	212	214	216	222	229
	17	211	217	219	222	225
	18	257	259	260	259	263
	19	229	234	238	239	245
	20	243	244	248	251	249
	21	259	264	264	266	272
	22	245	249	257	261	261
	23	277	278	277	277	277
	24	313	318	315	318	321
	25	343	343	354	358	362
	26	368	373	381	380	389
	27	383	385	384	397	403
	28	414	424	432	434	441
	29	454	464	469	470	474
	30	434	447	451	456	463
	31	484	490	495	503	513
	32	500	507	516	517	528
	33	506	513	516	525	534
	34	555	567	574	577	582
	35	530	538	551	551	559
	36	560	568	577	589	597
	37	630	650	657	678	701
	38	611	623	634	647	655
	39	728	744	757	780	799
	40	751	766	784	817	838
	41	755	770	796	815	834
	42	773	799	819	853	879

Month	Date of CDC Report				
	Jan. 87	July 87	Jan. 88	July 88	Jan. 89
43	889	906	936	965	984
44	954	976	1018	1057	1071
45	794	825	871	902	920
46	880	920	961	1018	1050
47	838	871	918	962	989
48	881	906	958	1014	1045
49	986	1059	1123	1197	1234
50	940	999	1072	1151	1191
51	966	1048	1135	1216	1263
52	990	1068	1145	1238	1278
53	1D35	1123	1241	1326	1382
54	1034	1164	1276	1396	1462
55	1017	1176	1331	1437	1497
56	982	1185	1308	1425	1494
57	894	1205	1343	1495	1577
58	761	1292	1468	1602	1706
59	373	1097	1265	1391	1464
60	43	1167	1397	1548	1627
61	.	1228	1499	1707	1811
62	.	1164	1483	1694	1815
63	.	1172	1548	1799	1946
64	.	874	1542	1816	1927
65	.	523	1564	1864	2012
66	.	81	1532	1873	2036
67	.	.	1508	1915	2094
68	.	.	1456	1905	2085
69	.	.	1374	1950	2130
70	.	.	1117	1933	2146
71	.	.	499	1776	1974
72	.	.	58	1877	2132
73	.	.	.	1765	2073
74	.	.	.	1654	2064
75	.	.	.	1763	2317
76	.	.	.	1296	2026
77	.	.	.	840	2061
78	.	.	.	125	2167
79	.	.	.	.	1959
80	.	.	.	.	1982
81	.	.	.	.	1656
82	.	.	.	.	1352
83	.	.	.	.	759
84	.	.	.	.	122

Table A. Five sets of data provided semi annually by the CDC beginning January 1987. These data are the reported numbers of diagnosed AIDS cases per month as of report time.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

SMOOTHING PROCEDURES FOR SIMULATED LONGITUDINAL MICRODATA

Jane F. Gentleman and Dale Robertson [1]

ABSTRACT

Microsimulation models allow one to study the behaviour of a large population over time. At Statistics Canada, health characteristics and risk factors are being added to a demographic and labour force model of the Canadian population. This paper describes a method for obtaining multivariate transition probabilities between states for use in advancing individuals in simulated time. The lack of longitudinal data means that these probabilities must be derived from cross-sectional data. The use of appropriate transition probabilities by the microsimulation model has the effect of producing smoother, more realistic, logically possible life histories. The probabilities are constrained to maintain consistency with the cross-sectional distributions. The constraints on the probabilities may be expressed as those of the transportation problem in network flow theory. The objective function in this special type of linear program is chosen to discourage unrealistically large or frequent changes of state across time. Canada Health Survey data were used to generate multivariate transition probability arrays for smoking, blood pressure, cholesterol, and body mass index, all thought to be important risk factors for coronary heart disease.

KEY WORDS: Longitudinal Data, Microsimulation, Simulation, Smoothing

INTRODUCTION

This paper describes techniques for enabling a dynamic microsimulation model which relies on cross-sectional source data to nevertheless produce realistically smooth simulated longitudinal microdata. A microsimulation model consists of a set of algorithms and a computer program which simulate microdata. The algorithms are based on probabilistic or deterministic submodels, and/or on observed distributions of real data. The microsimulation model generates a sample of simulated units which represent some conceptual population of units. These units might, for example, be people, households, or business firms. We shall refer to them as "individuals". The sample of individuals is used to make inferences about the population. Microsimulation models are particularly useful for posing and answering questions of a "what if" nature. To distinguish the data used in the construction of a microsimulation model from the data generated by such a model, we shall refer to the former as "source data" and to the latter as the "simulated data" or "sample data". For a useful, broad collection of papers concerning microsimulation, see Orcutt, Merz, and Quinke (1986).

A dynamic microsimulation model ages a sample of individuals across time, simulating multivariate data (such as marital status, employment status, education, consumption of manufactured goods, and health status) which describe them during each time period. There exist many panel and other surveys which can provide source data which are both multivariate and longitudinal, but the need of a microsimulation model for such data often cannot be fully met. Hoschka (1986, p. 49) lists "missing variables" and "cross section instead of panel surveys" as being among the most common shortcomings of microsimulation model source data. By their very nature, longitudinal data require a long period of time to be collected, and it is not always possible to foresee what combinations of variables will be needed, so that alternate strategies are needed.

Assume that for each variable of interest, a finite number of outcomes (or classes, or states) have been defined. From longitudinal age-specific source microdata, it is possible to estimate the distribution of a variable at a given age t, and to estimate transition probabilities for an individual moving from a certain state at age t to a certain state at age t+1. These probabilities can be used by the microsimulation model as it ages the sample.

[1] Analytical Studies Branch, Statistics Canada, Ottawa, Ontario K1A 0T6

In the absence of longitudinal source data, analysts often use cross-sectional data, treating the age-specific source data gathered at one point in time as if it were data describing one group of individuals across time. Transition probabilities cannot in general be deduced from cross-sectional data (a deficiency which also occurs with longitudinal data which are collected, but not linked, across time). However, if a microsimulation model ignores transitions and generates data independently for each age, the characteristics of a simulated individual may vary unrealistically across time, even though the distribution of the sample matches the source data distribution at each age.

For example, suppose that cross-sectional source data were used to estimate at each age the distribution of a variable describing an individual's smoking habit (classed as "Never Smoker", "Current Smoker", or "Former Smoker"). If the microsimulation model generates an individual's smoking habit independently for each age, the resulting simulated smoking history may have unrealistically frequent changes of state, and it may exhibit a logically impossible transition (such as from being a current smoker to being a never smoker).

Ideally, a microsimulation model would use an array of multivariate transition probabilities to move an individual from one age to the next. In the absence of multivariate source data, analysts may "synthetically" link different data files (enhancing the data for one individual by appending data from another, similar, individual), and they may resort to assuming independence of separate variables.

This paper describes procedures for obtaining multivariate transition probability arrays from cross-sectional multivariate data in order to smoothe the longitudinal behaviour of the simulated individuals. The examples provided utilize data for four variables having 5, 3, 3, and 4 states, respectively, so that there are 180 frequencies at each age group, and 180 x 180 = 32,400 transition probabilities from one age group to the next. Across the 12 age groups, there are therefore 356,400 transition probabilities. Given multivariate data for two adjacent age groups, an array of multivariate transition frequencies (and the corresponding array of transition probabilities) is obtained using linear programming (LP) methods. These transition frequencies are made consistent with the cross-sectional multivariate source data, and conditions which are innate to the particular variables are also imposed; these, plus the nonnegativity of the frequencies, form the constraints of the linear program. The linear program's objective function is chosen so that transitions to "nearby" states are favoured over transitions to "distant" states (which is reasonable if the time interval between the two age groups is relatively small).

The approach here is from a smoothing rather than an estimation point of view because of the very large number of degrees of freedom available for determining transition frequencies given relatively few marginal sums. Our approach is analogous to that used in smoothing ordinary univariate time series data, for which there are many possible smoothing algorithms; the choice of an algorithm and the parameter values thereof is often made heuristically, in order to obtain the desired quality and degree of smoothing. It is in that spirit that procedures are proposed here for generating realistically smooth longitudinal microdata.

THE SMOOTHING TECHNIQUE

Suppose that there are k variables of interest ($k \geq 1$), for which multinomial data are available as follows: For each variable, a finite set of mutually exclusive, exhaustive possible outcomes (states) has been defined, and cross-classified frequencies of occurrence of each outcome combination have been observed for n_t individuals of age t and for n_{t+1} individuals of age t+1. If the data are cross-sectional, these two groups of individuals are disjoint, and n_{t+1} may even be larger than n_t (which cannot occur in a closed population).

The array of transition frequencies between age t and age t+1 (and the corresponding array of transition probabilities) is 2k-dimensional. For notational simplicity, k will be assumed to be equal to 2, without loss of generality. Suppose, then, that the number of states for Var. 1 is s_1, and the number of states for Var. 2 is s_2. Let $u_{i_1 i_2}$ be the number of individuals who were observed to be in the bivariate state (i_1, i_2) at age t, and let

$v_{j_1j_2}$ be the observed number of individuals in state (j_1, j_2) at age t+1. (Here i_1 and j_1 label the state for Var. 1, and i_2 and j_2 label the state for Var. 2; i_1 and $j_1 = 1, \ldots, s_1$; i_2 and $j_2 = 1, \ldots, s_2$). Then $n_t = u_{\cdot\cdot}$ and $n_{t+1} = v_{\cdot\cdot}$ (where the dot notation signifies summation over the indicated subscript). Ordinarily, $n_t \neq n_{t+1}$; this occurs in a closed population because of losses due to mortality, and in cross-sectional data because the two groups contain different individuals.

For the time being, assume that there is no mortality between the two ages, and rescale the observed frequencies (for either or both ages) so that the number (n) of individuals represented at each age is the same: Multiply the $u_{i_1i_2}$'s by a constant C and the $v_{j_1j_2}$'s by $C\frac{n_t}{n_{t+1}}$. For example, multiply each observed frequency at age t by $C = \frac{n_{t+1}}{n_t}$, in which case the frequencies at age t+1 remain unchanged. Since the two sets of rescaled frequencies now have a common sum, the quantities $\left\{\frac{n_{t+1}}{n_t}u_{i_1i_2}\right\}$ and $\{v_{j_1j_2}\}$ can be treated as the marginal sums $\{x_{i_1i_2\cdot\cdot}\}$ and $\{x_{\cdot\cdot j_1j_2}\}$, respectively, of the array $\{x_{i_1i_2j_1j_2}\}$ of unknown transition frequencies which are to be determined using LP methods. As discussed below, the resulting transition probabilities are invariant to the choice of the scale factor C.

The transition frequency $x_{i_1i_2j_1j_2}$ is the unknown number of individuals who made the transition from state (i_1, i_2) at age t to state (j_1, j_2) at age t+1. The overall sum of the transition frequencies is $x_{\cdot\cdot\cdot\cdot} = Cn_t$ = n. The transition probability $p_{i_1i_2j_1j_2}$ is the probability of an individual being in state (j_1, j_2) at age t+1, conditional on having been in state (i_1, i_2) at age t:

$$p_{i_1i_2j_1j_2} = \frac{x_{i_1i_2j_1j_2}}{x_{i_1i_2\cdot\cdot}}.$$

(The word "probability" is used informally throughout the discussion here, and may instead be interpreted as "proportion".) The goal is to obtain reasonable values for the $p_{i_1i_2j_1j_2}$'s (or equivalently for the $x_{i_1i_2j_1j_2}$'s) for use in generating microsimulated data.

Using standard linear programming techniques, values $\{x_{i_1i_2j_1j_2}\}$ are determined so as to minimize an objective function, which is a weighted sum of the $x_{i_1i_2j_1j_2}$'s, subject to the following three types of constraints: (i) The frequencies must be non-negative; (ii) The marginal sums of the input multinomial data must be maintained; and (iii) Relationships innate to the variables must be maintained (e.g., that the number of transitions from being a current smoker to being a never smoker is zero).

The weights for the objective function are chosen here to favour stability by discouraging transitions to distant states, assuming that the concept of "distance" between states is meaningful. With the state labels suitably ordered, a reasonable choice for the weight $w_{i_1 i_2 j_1 j_2}$ for $x_{i_1 i_2 j_1 j_2}$ might be a measure of the distance between the state (i_1, i_2) at age t and the state (j_1, j_2) at age t+1, such as $| i_1 - j_1 | + | i_2 - j_2 |$ or $(i_1 - j_1)^2 + (i_2 - j_2)^2$.

There remains the question of which variables to use: the transition frequencies or the transition probabilities. That is, the result of minimizing

$$z = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} x_{i_1 i_2 j_1 j_2}$$

is in general different from the result of minimizing

$$z' = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} p_{i_1 i_2 j_1 j_2} = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w'_{i_1 i_2 j_1 j_2} x_{i_1 i_2 j_1 j_2}$$

(where $w'_{i_1 i_2 j_1 j_2} = \frac{w_{i_1 i_2 j_1 j_2}}{x_{i_1 i_2 \cdot \cdot}}$).

The LP method applied to the same observed data with two different choices of the rescaling factor C will yield the same array of transition probabilities (but not of transition frequencies) in both cases. If $\{x_{i_1 i_2 j_1 j_2}\}$ are the transition frequencies resulting from a choice of $C = C_1$, then it is straightforward to show that the transition frequencies resulting from an alternative choice of $C = C_2$ are $\left\{\frac{C_2}{C_1} x_{i_1 i_2 j_1 j_2}\right\}$. Both transition frequency arrays have the same transition probabilities.

It is instructive to interpret mortality as an additional variable for which two states - Alive and Dead - are defined at any given time. (Dead is an absorbing state; persons who are dead remain "forever" in the same multivariate state, and age for them is interpreted as the number of years since birth.) Viewed the other way around - as a life table to which variables representing other means of transition than dying have been added - the transition frequencies are similar to entries in a multistate life table (see, e.g., Rogers (1980)). In longitudinal data for a closed population, the numbers of individuals Alive and Dead are known at any given time, and transition frequencies between the two states are known; in cross-sectional data, the number Alive, but not the number Dead, is known, and no transition frequencies are known. It can be shown that rescaling of Alive individuals at age t is equivalent to removing from the Alive population at age t all of those individuals who are going to die by age t+1, but by applying the overall mortality rate rather than the state-specific rate. Since the LP method yields the same results regardless of the rescaling factor, this rescaling implies the use of a state-independent mortality rate.

This is what occurs when cross-sectional data are rescaled; the resulting transition probabilities may be interpreted as transition probabilities from age t to age t+1 for only those individuals who survived to age t+1, but under the assumption that all state-specific mortality rates are the same. Using rescaled cross-sectional data is not ideal in that one likely reason for defining different states of a variable is that the mortality rate is thought to be state-dependent. Intuitively speaking, a larger percentage of high risk individuals should die in [t,t+1), resulting in relatively fewer of them alive at age t+1. Assuming a uniform mortality rate for all states makes it appear that some high risk individuals have moved to lower risk states.

EXAMPLES

The input data used to demonstrate the smoothing technique are from the 1978/79 Canada Health Survey (CHS). This was a multistage stratified household survey of 31,668 individuals. For details of the CHS, see Statistics Canada and National Health and Welfare (1981). For each sex, and for each of 12 age groups (15-19, 20-24, 25-29, ..., 65-69, and 70+), cross-classified frequencies were obtained for the following variables and classes:

Var. 1: Body Mass Index $\left(\frac{kg}{m^2}\right)$

(i) <20
(ii) [20,25]
(iii) (25,27]
(iv) (27,30]
(v) >30

Var. 2: Serum Cholesterol $\left(\frac{mg}{dL}\right)$

(i) ≤200
(ii) (200,240]
(iii) >240

Var. 3: Diastolic Blood Pressure (mmHg)

(i) <90
(ii) [90,105)
(iii) ≥105

Var. 4: Smoking Habit

(i) Never Smoker
(ii) 1-20 cigarettes/day
(iii) >20 cigarettes/day
(iv) Former Smoker

Frequencies were calculated using the survey weights. These four variables are risk factors which can be used to help predict coronary heart disease. The transition probabilities derived here are to be used in a health microsimulation model being developed by Wolfson (1989); a submodel, constructed by Wolfson and Birkett (1989), simulates the onset and progression of coronary heart disease.

Transitions involving the smoking variable have certain innate constraints. The probability of becoming a former smoker immediately after being a never smoker is zero (for a short age increment during which it is assumed that only one transition occurs). The probability of becoming a never smoker after being in any of the other three smoking categories is zero. And it may be reasonable to assume that the probability of quitting smoking is less than or equal to the probability of resuming smoking. The appropriate elements of the transition probability array must therefore obey certain equations or inequalities. The LP approach can maintain such relationships by imposing them as additional constraints.

Table 1 gives a one-variable example of the LP procedure results. Transition frequencies and probabilities for males from age group 30-34 to age group 35-39 were calculated using the observed marginal distributions of just the Smoking Habit variable. The four smoking states were ordered as they might occur for one individual across time - from never smoker to lighter smoker (1-20 cigarettes) to heavier smoker (>20 cigarettes) to former smoker. The LP procedure was applied using different combinations of weights (absolute distance or squared distance) and objective functions (z or z'). In all four cases, the (1,4), (2,1), (3,1), and (4,1) elements of the transition matrices (involving transitions from never smoker to former smoker, and from lighter smoker, heavier smoker, and former smoker to never smoker) were constrained to be zero, but no inequality constraints were imposed for quitting smoking relative to resuming smoking.

In Table 1, changing from z to z' made a difference when absolute distance weights were used, but made no difference when squared distance weights were used. In general, the use of absolute distance weights permits transitions to more distant states to occur than with squared weights (or weights based on a more rapidly increasing function of distance). In the four examples, the only transitions permitted over a distance of more than one state are from being a lighter smoker to being a former smoker (absolute distance weights, objective function z, probability .08), and from never having smoked to being a heavier smoker (absolute distance weights, objective function z', probability .02). In the latter case, however, the one-state move from never having smoked to being a lighter smoker is less probable than the two-state move (in fact, the one-state transition is impossible), which may be unrealistic.

On the other hand, absolute distance weights generally result in larger diagonals. The diagonals of the examples using absolute distance weights are greater than or equal to the corresponding diagonals of those using squared distance weights. All four examples have probabilities of 1.00, which is probably unrealistically large, for the zero-state transition from former smoker to former smoker. Even so, this does not imply that a quitter of smoking will remain a quitter forever, as each age transition uses a different set of transition probabilities.

The probability of resuming smoking (the sum of elements (4,2) and (4,3)) is zero in all four examples, and the probability of quitting (the sum of elements (2,4) and (3,4)) is higher - either .08 or .07. Attempts to force the probability of quitting to be lower than the probability of resuming resulted in the LP program halting in some cases because no feasible solution exists for these data under these constraints. The problem is caused by the use of cross-sectional data and by incorrect assumptions about them, not by the LP method.

It is useful to inspect the transition probabilities and to examine the effects of varying the parameters, as one does when smoothing time series. One can examine the various trade-offs among the different solutions and select the most appropriate one for the microsimulation model. The acceptability of a set of transition probabilities depends strongly on the particular set of data and on the assumptions. For example, our assumption here that transitions from never smoker to former smoker are impossible is perhaps overly stringent for five-year age intervals.

Figure 1 shows two synthetic life histories simulated using the multivariate distributions for the four variables across all 12 age groups. In Figure 1A, the 11 8-dimensional transition probability arrays obtained using objective function z and squared distance weights were used to assign states to one individual (named "Sam" for "smooth"). In Figure 1B, only the distributions at each age were used, so that the multivariate state of this individual (named "Roy" for "rough") at age t is independent of his state at age t+1. The greater smoothness and continuity of Sam's life history is clearly noticeable. Only once does he jump from one state across an intermediate state to another state, while such leaps are quite frequent in Roy's life. Note that Roy twice experiences a forbidden transition, becoming a never smoker at age 50-54, and then becoming a former smoker at age 55-59. Roy's body mass index fluctuates unrealistically, as does his cholesterol. Sam's body mass index follows quite a believable pattern of increase to age 50-54 and then declines, similar to his cholesterol. In smoking, he shows a plausible pattern, in general gradually increasing his consumption until he quits in middle age. Roy's smoking pattern is logically impossible. Both of these individuals have reasonable blood pressure patterns.

COMPUTATIONAL CONSIDERATIONS

The procedure used to obtain the transition frequencies has been described above as a linear programming one. In fact, the problem falls into a very important special class of linear programs, i.e., network flow problems, and a special case of these known as the transportation problem. The term transportation problem arose from the original interpretation as finding the least costly way to route materials from supply points to demand points (see Hitchcock (1941)). The essential constraints of a transportation problem are the imposition of known row and column totals on the non-negative elements of a matrix. By choosing some convenient ordering for the multivariate states, one can imagine the transition frequencies to be matrix elements, with the row labels corresponding to the starting states and the column labels to the ending states. The number of individuals making a transition is clearly non-negative, and the row and column sums correspond to the total numbers of individuals in the initial and final states. We want to impose consistency with these given totals, which is precisely the transportation problem framework. The objective function in our problem imposes costs on various transitions. These are analogous to the shipping costs from one place to another in the classical application. The coefficients are such that transitions to "nearer" states are cheaper than those to more distant states.

The recognition that our linear programming problem is a network flow problem has important theoretical and practical consequences. One pleasant property is that integer valued solutions are found. If the row and column totals are integers, the algorithms will return optimal solutions with integer values, so that the number of individuals making a transition is never fractional (see Lawler (1976)). Finding integer valued optimal solutions is difficult for a general linear program, but it is automatic with a network flow.

Other advantages of phrasing our problem as a network flow are that network flow problems can be solved orders of magnitude faster and with less computer storage space than general linear programs. The problems solved here with less than 400 nodes are not large ones by the standards of the field and are routinely solved by available codes. In one large application (see Barr and Turner (1981)), a transportation problem with more than 20,000 constraints and 10,000,000 variables was solved.

In the simplest transportation problem, all transitions are allowed, with no upper limit on the value of an individual frequency. In this case, providing that the row and column totals are consistent (i.e., have the same overall total), the problem always has a solution (i.e., is "feasible", in the terminology of mathematical programming). The framework does have more flexibility than this simple form implies, and this flexibility is needed for our problem. For example, if certain transitions are logically impossible, then they can be forbidden. One may also impose lower and upper bounds on various variables in the solution. This corresponds to restricting the ranges of certain transition probabilities to reflect knowledge and beliefs about what is likely. When additional constraints of this type are added, the problem may no longer be feasible. (As a simple example, if enough transitions are forbidden, it may not be possible to satisfy the demand at a particular node or nodes.) In our data, infeasibility was encountered in some cases and was always traced to the same cause: our constraints made it impossible to become a never smoker from any other state (which was reasonable), but the raw data, after re-scaling to achieve consistent overall sums, had more never smokers at age t+1 than at age t. The resulting infeasibility was not due to a problem with the method; it was a consequence of the use of cross-sectional data as a substitute for longitudinal data. In a sense, the re-scaled data contain outliers (values inconsistent with the model) and must be adjusted to meet the logically necessary condition that the proportion of never smokers can only stay the same or decrease.

Our results were obtained using two SAS Procedures: LP for general linear programs (SAS Institute (1985)), and NETFLOW for network flow problems (SAS Institute (1985)). Another Procedure, TRANS (SAS Institute (1985), specializes in transportation problems, but internal difficulties in our version, since rectified by SAS, caused us to abandon its use. TNETFLOW, a superior Procedure for network flow problems (SAS Institute (1986)), has now been produced.

Figure 2A shows the observed percentages of never smokers (males only) in our Canada Health Survey data across the 12 age groups. Except for one large percentage at age 15-19 and one small one at age 50-54, the percentage across the ages is roughly linear with a negative slope, but it is not uniformly non-increasing. To adjust these data, we fitted a simple linear regression on age of the logarithm of the proportion of never smokers, omitting the above-mentioned two points. Fitted values on this line were substituted for observed data when necessary to obtain a non-increasing proportion of never smokers. The relative magnitudes of the proportions of individuals in the other smoking categories were maintained.

Figure 2B shows the observed percentages of female never smokers across the 12 age groups. These show a clearly increasing trend over time, very likely due to a cohort effect in this cross-sectional data: older women in the 1978 population were more likely to be never smokers than were younger women. This illustrates a severe conflict caused by the use of cross-sectional data in place of longitudinal data. If such data were casually subjected to the procedures described above, no solutions would be found for most age transitions. The smoothing procedure thus has a side benefit of providing a warning about certain types of inconsistencies in the input data.

CONCLUDING REMARKS

The statistical concepts of heterogeneity, selection, and multistate life tables all come together when considering the problems introduced by using cross-sectional data in place of longitudinal data. More research is needed to determine how to recognize and correct for these problems, and more longitudinal data are needed in order to sidestep them.

On the computing front, despite what may sometimes be insurmountably high requirements of microsimulation models for resources, there is optimism that further technological advances - such as increased processing speed, higher capacity data storage, more use of dedicated computers, and parallel processing - will allow microsimulation models to expand and improve their capabilities (see Hoschka (1986)).

The authors wish to thank Michael Wolfson for motivating this study, the results of which will be implemented in his POHEM health microsimulation model (see Wolfson (1989)). We also thank Monica Tomiak for her invaluable technical assistance and programming support.

REFERENCES

Barr, R.S. and Turner, J.S. (1981). Microdata File Merging Through Large Scale Network Technology. Math. Prog. Study, **15**, 1-22.

Hitchcock, F.L. (1941). The Distribution of a Product from Several Sources to Numerous Localities. J. Mathematical Physics, **20**, 224-230.

Hoschka, Peter (1986). Requisite Research on Methods and Tools for Microanalytic Simulation Models. In Orcutt, et al. (1986), pp. 45-54.

Lawler, Eugene L. (1976). **Combinatorial Optimization: Networks and Matroids**. Holt Rinehart and Winston.

Orcutt, Guy; Merz, Joachim; and Quinke, Hermann (1986), Editors. **Microanalytic Simulation Models to Support Social and Financial Policy**. Proceedings of 1983 symposium in Bonn, Germany. North-Holland.

Rogers, Andrei (editor) (1980). Essays in Multistate Mathematical Demography. Special issue of Environment and Planning A 12(5).

SAS Institute Inc. (1985). **SAS/OR User's Guide**, Version 5 Edition.

SAS Institute Inc. (1986). Technical Report: P-146. Changes and Enhancements to the Version 5 SAS System.

Statistics Canada and National Health and Welfare (1981). The Health of Canadians. Report of the Canada Health Survey. Catalogue 82-538E. Ottawa. Canada.

Wolfson, Michael C. (1989). A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data. Paper presented at 21st General Conference of the International Association for Research in Income and Wealth, Lahnstein, West Germany, Aug. 20-26, 1989.

Wolfson, Michael and Birkett, Nick (1989). POHEM, Population Health Module of the System of Health Statistics (SHS): Preliminary Exploration of CHD. Paper presented at Canadian Epidemiology Research Conference, Ottawa, August 1989.

TABLE 1

A. TRANSITION FREQUENCIES FOR ONE VARIABLE CASE
(VAR. 4 : SMOKING HABIT) FOR MALES FROM AGE GROUP 30-34 TO AGE GROUP 35-39
ELEMENTS (1,4), (2,1), (3,1), AND (4,1) ARE CONSTRAINED TO BE ZERO

$W_{ij} = |i-j|$

		NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER	
	NEVER SMOKER	152,389	3,263	0	0	155,652
Z	1-20 CIG/DAY	0	111,080	24,390	11,661	147,131
	)20 CIG/DAY	0	0	161,514	0	161,514
	FORMER SMOKER	0	0	0	222,417	222,417
		152,389	114,343	185,904	234,078	686,714

$W_{ij} = (i-j)^2$

	NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER	
	152,389	3,263	0	0	155,652
Z	0	111,080	36,051	0	147,131
	0	0	149,853	11,661	161,514
	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

$W_{ij} = |i-j|$

		NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER	
	NEVER SMOKER	152,389	0	3,263	0	155,652
Z'	1-20 CIG/DAY	0	114,343	32,788	0	147,131
	)20 CIG/DAY	0	0	149,853	11,661	161,514
	FORMER SMOKER	0	0	0	222,417	222,417
		152,389	114,343	185,904	234,078	686,714

$W_{ij} = (i-j)^2$

	NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER	
	152,389	3,263	0	0	155,652
Z'	0	111,080	36,051	D	147,131
	0	0	149,853	11,661	161,514
	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

B. TRANSITION PROBABILITIES FOR ONE VARIABLE CASE
(VAR. 4 : SMOKING HABIT) FOR MALES FROM AGE GROUP 30-34 TO AGE GROUP 35-39
ELEMENTS (1,4), (2,1), (3,1), AND (4,1) ARE CONSTRAINED TO BE ZERO

$W_{ij} = |i-j|$

		NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER
	NEVER SMOKER	.98	.02	.00	.00
Z	1-20 CIG/DAY	.00	.75	.17	.08
	)20 CIG/DAY	.00	.00	1.00	.00
	FORMER SMOKER	.00	.00	.00	1.00

$W_{ij} = (i-j)^2$

	NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER
	.98	.02	.00	.00
Z	.00	.75	.25	.00
	.00	.00	.93	.07
	.00	.00	.00	1.00

$W_{ij} = |i-j|$

		NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER
	NEVER SMOKER	.98	.00	.02	.00
Z'	1-20 CIG/DAY	.00	.78	.22	.00
	)20 CIG/DAY	.00	.00	.93	.07
	FORMER SMOKER	.00	.00	.00	1.00

$W_{ij} = (i-j)^2$

	NEVER SMOKER	1-20 CIG/DAY	)20 CIG/DAY	FORMER SMOKER
	.98	.02	.00	.00
Z'	.00	.75	.25	.00
	.00	.00	.93	.07
	.00	.00	.00	1.00

FIGURE 1. TWO SIMULATED LIFE HISTORIES

A. SAM (Smooth history)

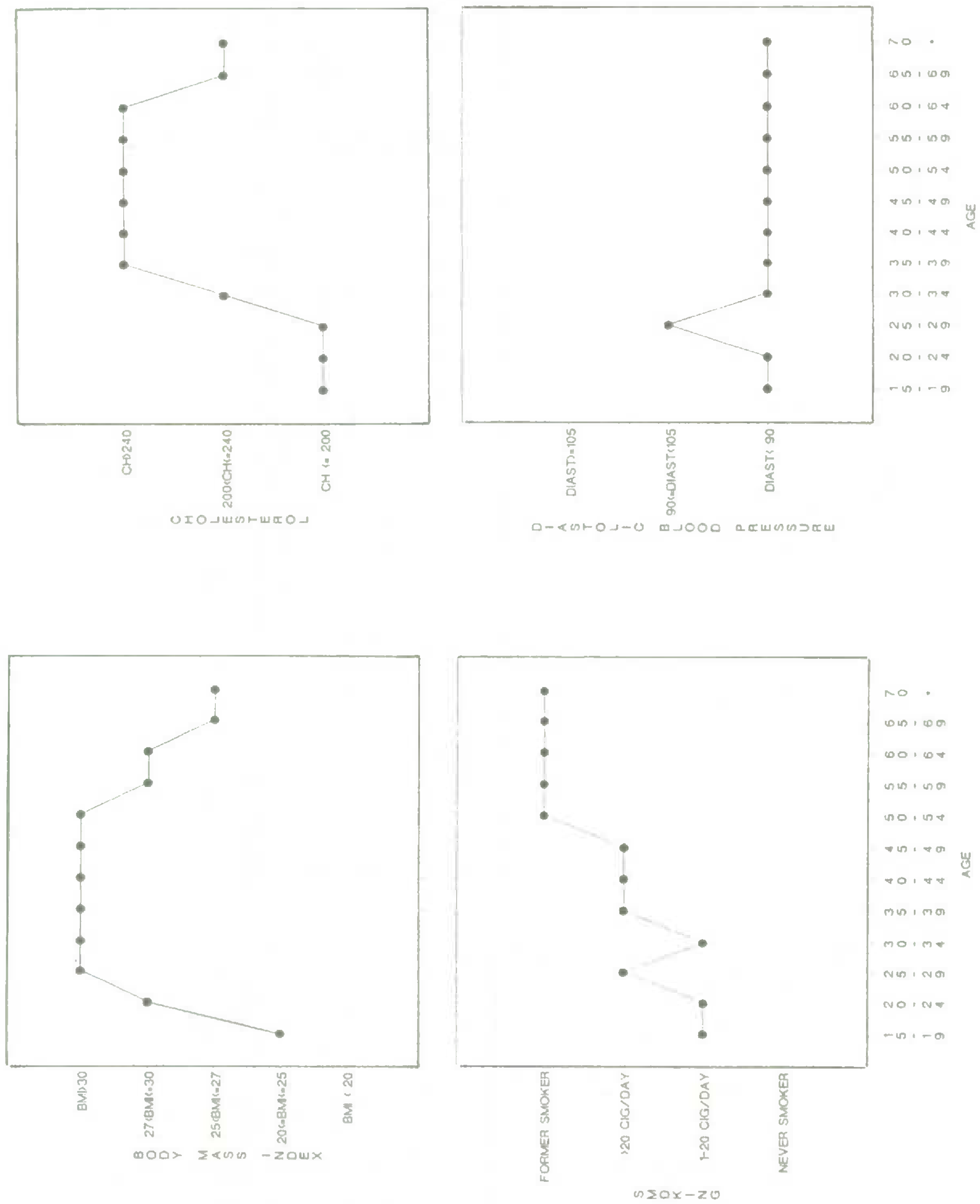


FIGURE 1 (continued).

B. ROY (Rough history)

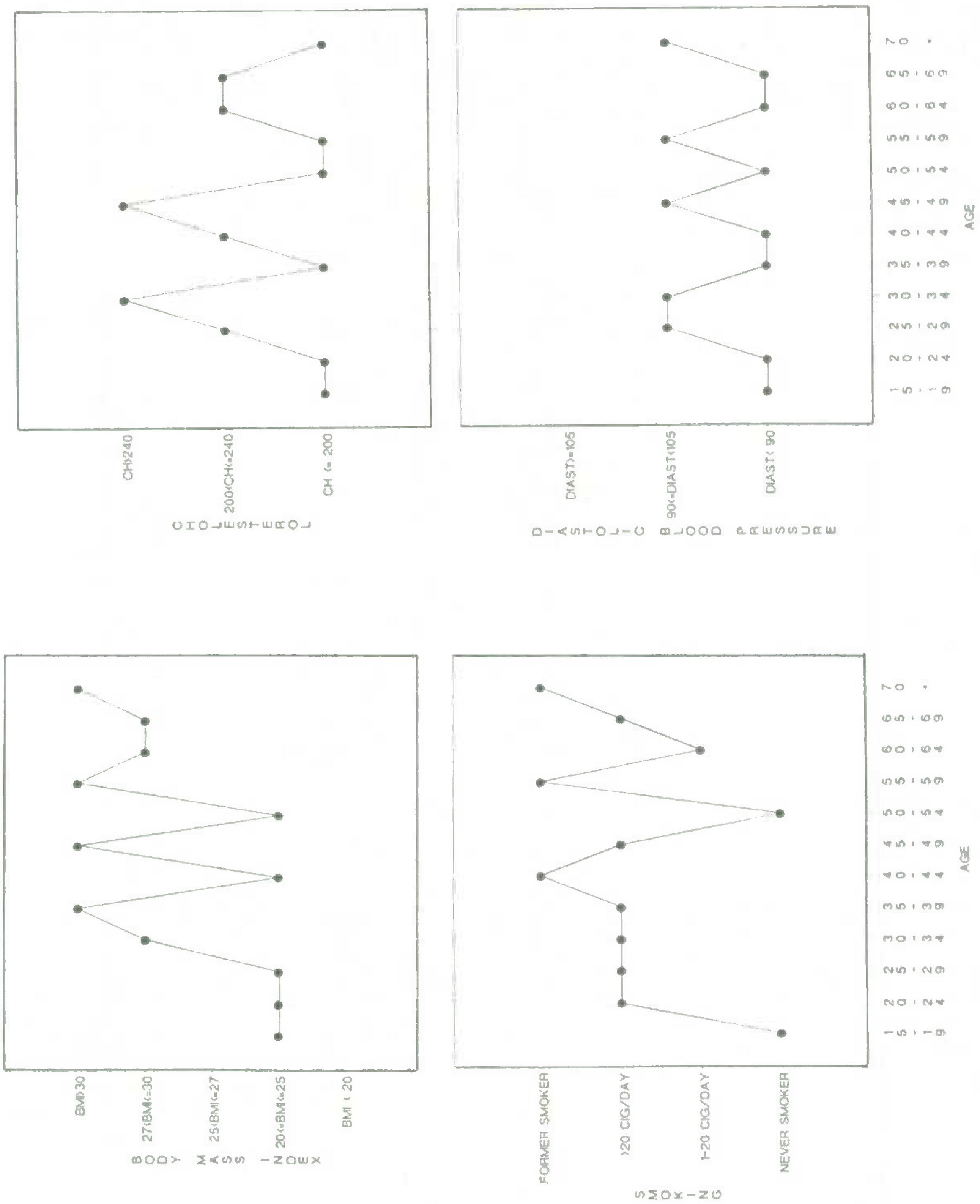
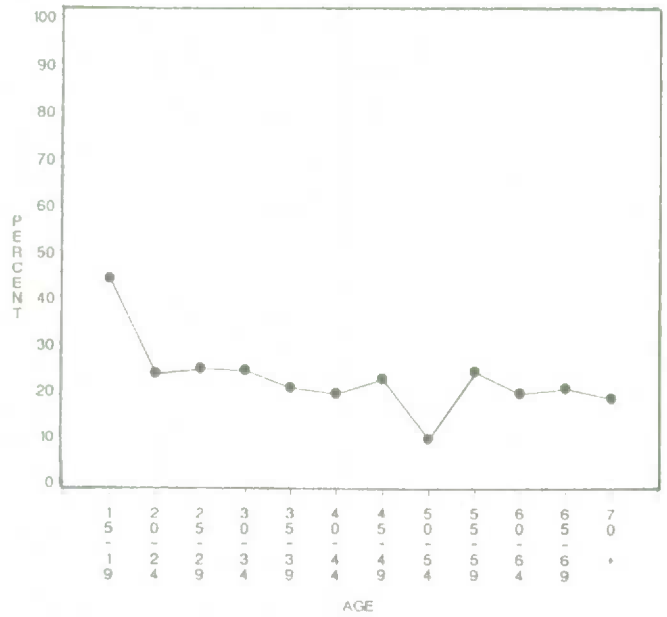
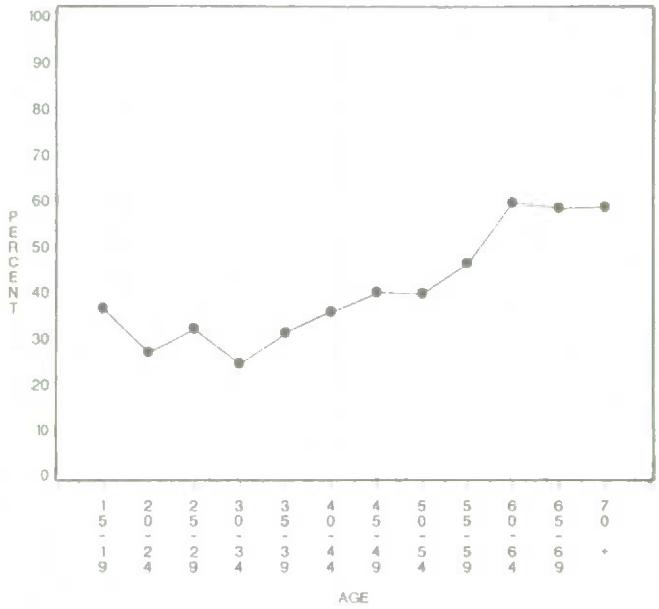


FIGURE 2. PERCENTAGES OF NEVER SMOKERS ACROSS 12 AGE GROUPS

A. MALES



B. FEMALES



Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

CHANGING DEMOGRAPHIC CHARACTERISTICS AND CARDIOVASCULAR DISEASE

H. Johansen[1], C. Nair[2], M. Nargundkar[3], and J. Strachan[4]

ABSTRACT

Key Words: Cardiovascular disease, mortality rates,immigrants, ethnicity, place of birth, time trends.

Cardiovascular disease (CVD) is the major cause of death in Canada as it is in most industrialized countries. Major risk factors for CVD have been identified as smoking, high blood pressure and cholesterol. Approximately one in six Canadian residents is a first generation immigrant. It has been shown that the CVD mortality rates vary by ethnicity and were found to differ for the first generation ethnic groups in Canada. Overall lower CVD mortality rates were found for first generation Canadians from South America, China and South Asia; high rates are indicated for Scandinavia and Africa. The rates for North America are similar to those found for Eastern and Western Europe. Between two five year time periods (1969-73 and 1984-88) CVD mortality rates generally were found to decrease except for Africa (age 35+). Similar ranking of CVD mortality rates were found by age and sex. The rates were consistently higher for males than for females.

1. INTRODUCTION

Cardiovascular disease (CVD) is the major cause of death in Canada as it is in most industrialized countries. It is the leading cause of hospitalization in Canada with direct hospital costs estimated to be in excess of 3 billion dollars per year (Nair et al., 1989). During 1988, 77,000 Canadians died from CVD (Nair et al., 1989). This represents a group of people the size of the population of Kingston, Ontario. CVD has been the leading cause of death in Canada over a long period of time. In 1921, deaths due to CVD were 18.6% of all deaths; in 1981 CVD is still the leading cause of deaths and accounts for 46.6% (Figure 1) of all deaths.

Leading Causes of Death,
Canada, 1921 and 1981

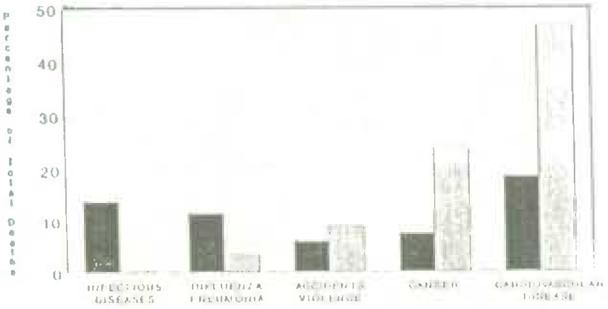


Figure 1

In 1987, 43% of all deaths were due to CVD. This implies that 4 out of 10 Canadians are likely to die from CVD each year. Figure 2 shows the death rates from CVD during 1987 by age and sex. It shows that death from CVD starts at an earlier age for males and for all ages the rates for males are higher than for females. As is to be expected, persons over the age of 65 have substantially higher death rates than the younger age groups. The death rates for persons of age 75 years and over are approximately four times that for persons less than 65 years. This has major implications for health services in an aging population. Figure 3 gives the age standardized mortality rates (ASMR's) by sex from 1951 to 1987. The good news is that the ASMR's have declined in the last thirty five years. If the rate had remained at its highest level there would have been 22,000 additional male deaths and 13,000 additional female deaths in 1987.

Three major risk factors for CVD have been identified: smoking, high blood pressure, and elevated serum cholesterol. Comparable data for two of these risk factors, namely smoking and high blood pressure, are

[1] H. Johansen, Health Promotion Directorate, Health and Welfare Canada, Ottawa, Ontario K1A 1B4.

[2] C. Nair, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario K1A 1B4.

[3] M. Nargundkar, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 1B4.

[4] J. Strachan, School of Health Information Science, University of Victoria, Victoria, B.C. V8W 2Y2.

Cardiovascular Disease
Rate per 100,000 Population
Canada, 1987

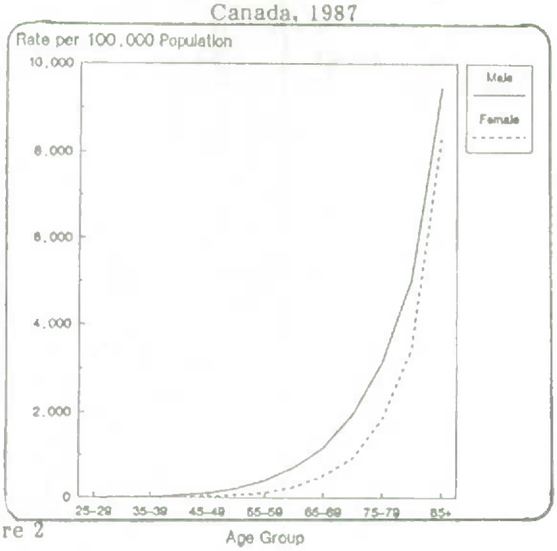


Figure 2

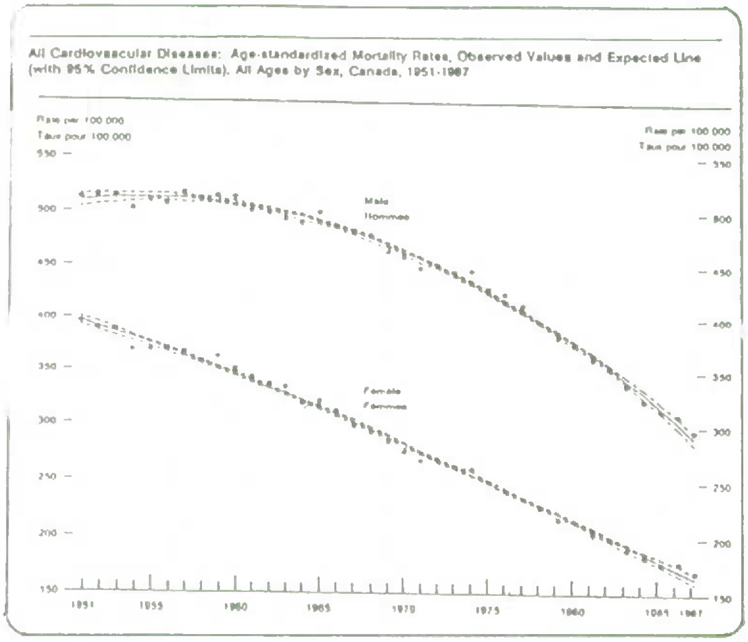

All Cardiovascular Diseases: Age-standardized Mortality Rates, Observed Values and Expected Line (with 95% Confidence Limits), All Ages by Sex, Canada, 1951-1987

Figure 3

available over a period of time and are presented in this paper. The proportion of smokers in Canada has declined from 42.2% in 1970 to 29.4% in 1986, (Figure 4), (Health & Welfare, 1973). Similarly, the proportion of Canadians with diastolic blood pressure >= 100 mmHg has decreased from 9.6% in the Nutrition Canada Survey (NCS), (Health & Welfare, 1973) to 4.9% in the Canadian Health Survey (CHS), (Health & Welfare, 1981) to 2.7% in the Canadian Blood Pressure Survey (CBPS), (Figure 4), (Health & Welfare, 1989). This may be part of the reason for the decline in ASMR's over the past three decades.

Currently, approximately one in six Canadian residents is a first generation immigrant (Employment & Immigration, 1988). In recent years, the trend of immigration has shifted from European countries to countries from other parts of the world. Figure 5 shows this shift by the region of birth of immigrants over a period of time. Immigration from Europe has decreased while immigration from Asia and South America has increased. This trend has contributed to the multi-cultural and multi-ethnic mosaic of the Canadian population. It has been shown that CVD rates vary by region (Thom, 1988). Therefore epidemiological studies of first generation Canadians by place of birth may be used to help evaluate the health status of Canada and help in the planning of health intervention programs. This paper will examine the CVD mortality rates of immigrants to Canada by region of birth.

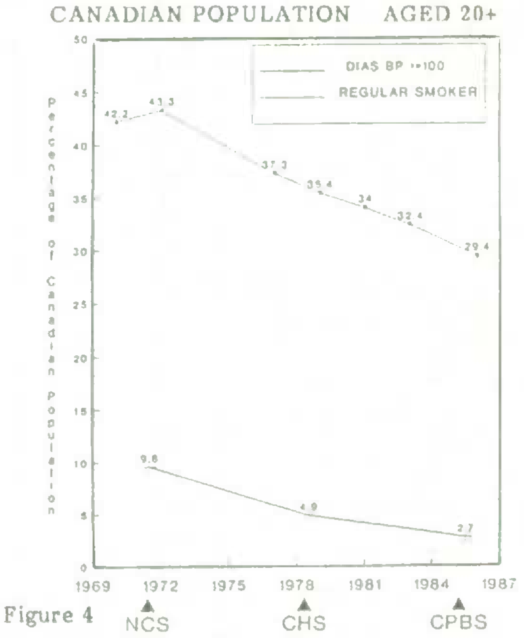


Figure 4

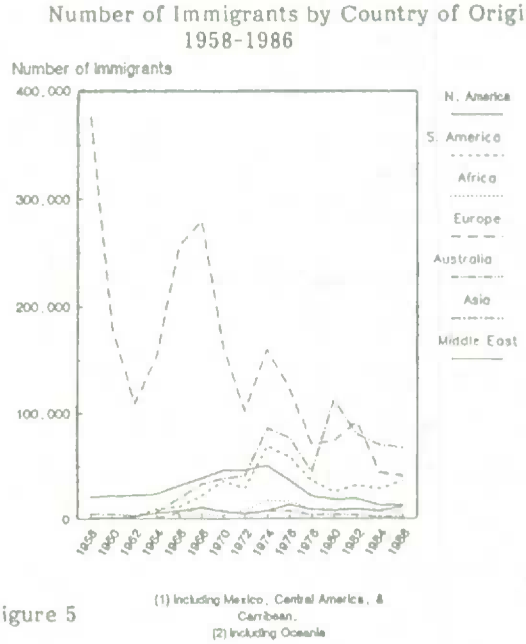


Figure 5

2. METHODOLOGY

The data used in this study are from the Canadian Census of Population (Statistics Canada, 1971 & 1986) and the Canadian Mortality Data Base (CMDB) (Statistics Canada). For the purposes of the study and to act as a proxy for ethnic differences, countries of birth were grouped geographically as outlined in Table 1. The number of countries selected for the study was limited by the availability of data on the CMDB. For example, Japan and Australia were ultimately excluded from most of the analysis due to small numbers. Time periods for analysis were also limited by the availability of data, (Table 2). The time periods 1969-73 and 1984-88 were selected since the data was the most complete during these two five year time periods.

In Alberta for 1984 the total number of deaths due to CVD was known but the country of birth was not recorded and therefore had to be estimated. The number of deaths in Canada for 1988 were estimated from the number of deaths in 1987. All estimates were evaluated by checking their consistency with known data trends.

Five year time periods were chosen to provide a reasonable number of deaths in the selected groups of countries. The 1971 and 1986 census data - corresponding to the mid-points of the five year time periods - were categorized by country of birth, age (35-54, 55-64, 65+) and sex. The data on the CMDB were similarly grouped for age, sex, and country of birth as well as by cause of death (all causes, CVD, Ischemic Heart Disease (IHD), and Cerebrovascular Disease (CBVD)). The census data were used for the denominator and the sum of the deaths over each 5 year period was used in the numerator giving the five year grouped mortality rates. Age standardized mortality rates (ASMR's) were calculated using the 1986 Canadian population as the standard. When ASMR's are referred to in the following text it refers to 5 year grouped data, age standardized in the above mentioned three age groups starting at age 35.

3. RESULTS

Figure 6 shows the ratio of CVD to total mortality in Canada, by country of birth, for the two time periods: 1969-73, 1984-88 respectively. In 1969-73 CVD accounted for at least fifty percent of all deaths over age 35. In 1984-88, the percentages of deaths due to CVD has declined in most countries of birth to between 40-50%.

The Canadian ASMR's by country of birth for death from all causes for the two time periods are shown in Figure 7. All cause mortality rates vary by country of birth in 1984-88, the lowest being for South America and South Asia and the highest for Africa. ASMR's decreased between 1969-73 and 1984-88 for all regions except Africa.

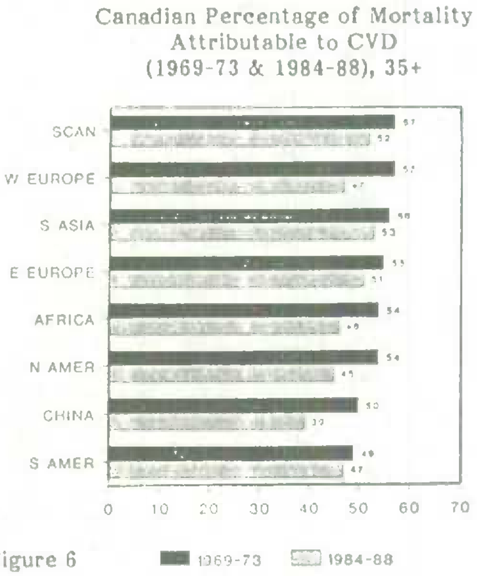


Figure 6

Figure 8 gives the Canadian ASMR's by country of birth for CVD mortality for age 35+. CVD ASMR's also vary by country of birth and show a similar pattern as death from all causes. Rates are decreasing except for Africa for ages 35+, with a major decline indicated for China. The two time periods have a similar ranking by country of birth.

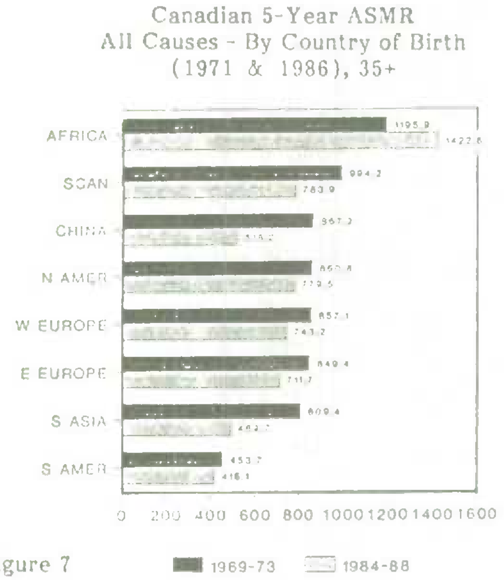


Figure 7

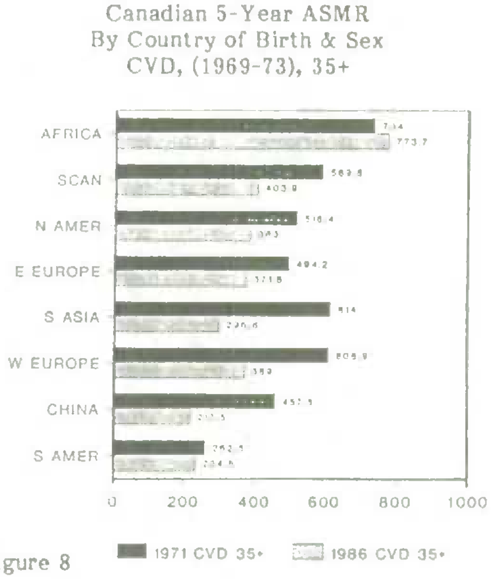


Figure 8

Male/female CVD differences for the two time periods are given in Figures 9 & 10. Males consistently had a higher rate than females for all countries. Chinese women had substantially lower rates than Chinese males. Males and females had a somewhat similar ranking by country of birth.

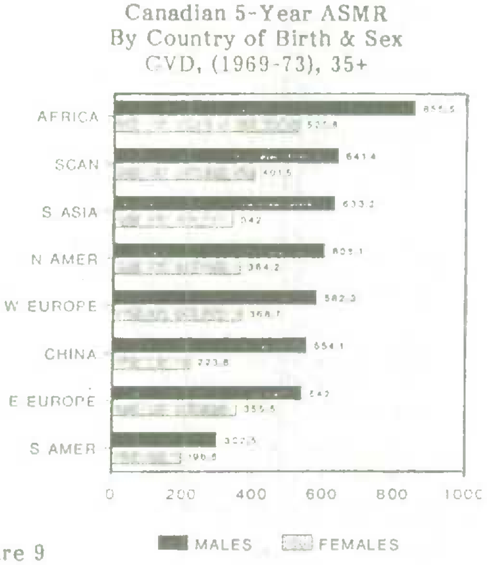


Figure 9

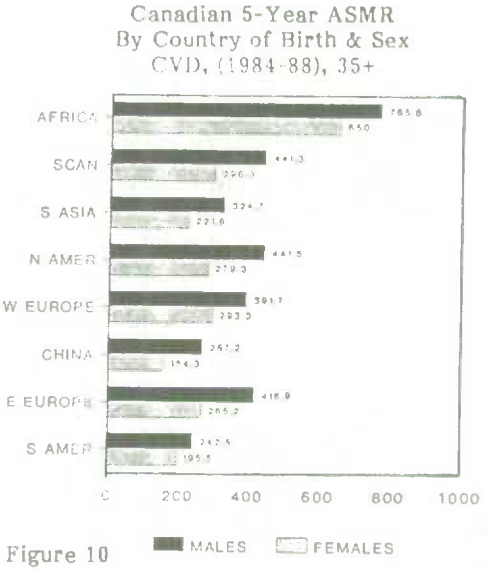


Figure 10

Figures 11 & 12 give the Canadian 5-year grouped age specific CVD mortality rates by country of birth for 1984-88 for males and females. Rates increased steeply by age. Roughly the same pattern of high and low rates by birth country is followed by age and sex.

Figure 13 gives the Canadian ASMR's for Ischemic Heart Disease (IHD), Cerebrovascular disease (CBVD), and for other CVD by country of birth for 1984-88. IHD accounts for over 50% of all deaths from CVD. CBVD accounts for much less but still makes a significant contribution. Note that CBVD accounts for 1/3 of all CVD deaths for those born in China.

Figures 14 & 15 show Canadian ASMR's for IHD and CBVD for 1969-73 and 1984-88. IHD declined in China and South Asia. There is roughly a similar pattern for IHD by country of birth.

5-Yr Age Specific CVD Mortality
By Country of Birth
1984-1988

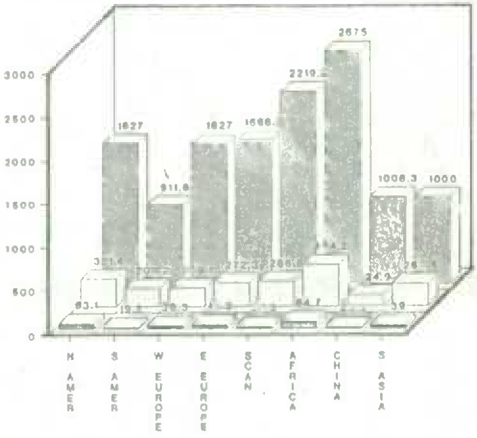


Figure 11

5-Yr Age Specific CVD Mortality
By Country of Birth
1984-1988

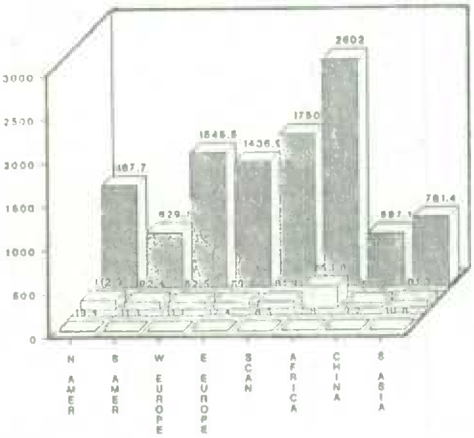


Figure 12

Canadian 5-Year ASMR
IHD, CBVD & Other CVD
(1984-88), 35+

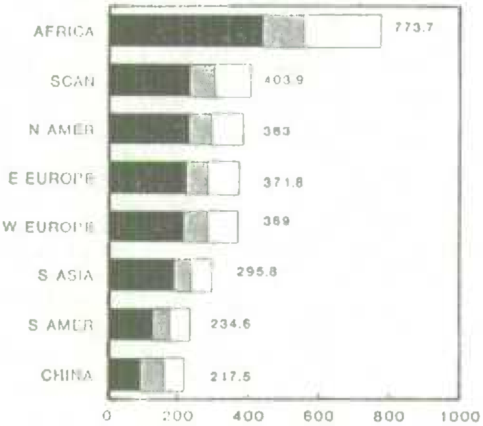


Figure 13

Canadian 5-Year ASMR
IHD, CBVD, 35+
(1969-73)

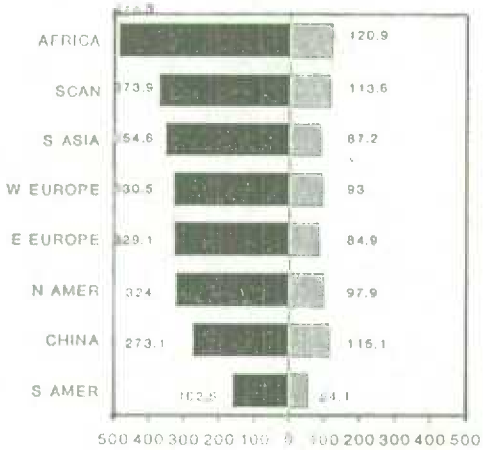


Figure 14

Canadian 5-Year ASMR
IHD, CBVD, 35+
(1984-88)

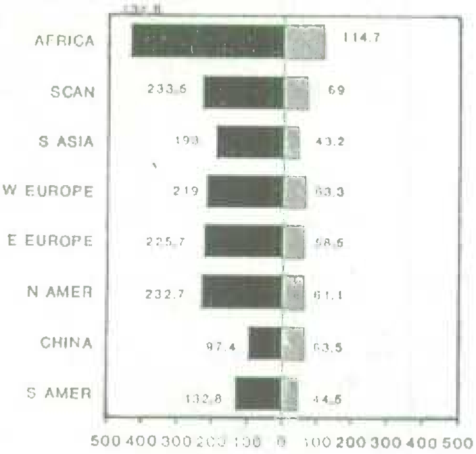


Figure 15

Figure 16 transposes the CVD rates for Canadians on a map of the world. Overall, low rates were found for first generation Canadians for South America, China, and South Asia; high rates are indicated for Scandinavia and Africa; North American rates are similar to those found in Eastern and Western Europe. We compared tHis general ranking of CVD mortality rates for Canadian immigrants to mortality rates found by WHO (World Health Organization, 1988) for their countries of birth (Tables 4 and 5). As far as can be determined there is general agreement with the above ordering. Due to a lack of detailed information rigorous comparisons of ASMR's of first generation Canadians with the ASMR's of their country of origin were not possible.

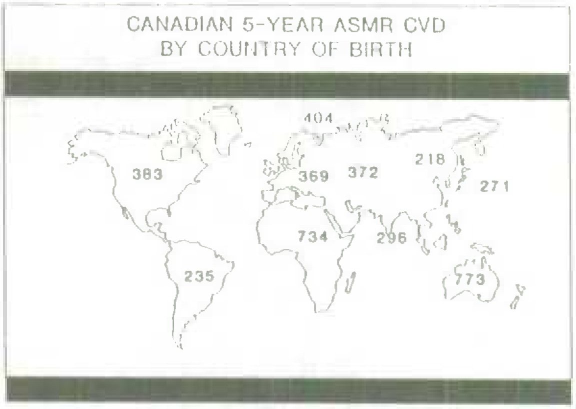


Figure 16

4. DISCUSSION

The mortality due to CVD has decreased over the past thirty-five years, but it is still the major cause of death in Canada. Over 40% of all deaths are attributable to CVD. Males tend to have higher CVD mortality rates than females. This similar pattern holds for the 5-year ASMR's for first generation Canadians. However, the magnitude of the 5-year ASMR's for first generation Canadians vary significantly by country of birth. The rates are high for Scandinavia and Africa and low for South America, China, and South Asia. The rates are decreasing for the general population and for first generation Canadians except those from Africa for ages 35+. WHO reports that CVD rates are decreasing worldwide except for Eastern European countries (World Health Organization, 1988). We did not find an increase for Eastern European Canadian immigrants. This might be explained by the selection process for immigration and a different diet. Comparable data for African countries was not available. However, a recent paper by Balfe et al., 1988 demonstrated that CVD rates are very high for South African whites and Asians. These two groups constitute the majority of the first generation Canadians born in Africa.

A recently published review article by McKeigue et al., 1989 on mortality due to CVD of South Asians living in Asian and non-Asian countries, indicates the CVD rates for South Asians are among the highest in the world. The article also discusses the South Asian populations predisposition to diabetes mellitus and particularly to non-insulin dependent diabetes mellitus. The relationship between these two diseases does seem to exist (Salonen, 1989; Hughes et. al., 1989a; Hughes et. al., 1989b) but it is unknown if the relationship is causal or genetic. The high CVD rates for South Asians found by McKeigue et al., 1989 and various others (Hughes et. al., 1989a; Hughes et. al., 1989b) is not consistent with the findings from our study on mortality rates in first generation Canadians from South Asia.

Our study indicates that the ASMR's for CVD and all causes has been consistently lower than many other Canadian immigrant groups studied. However, it also shows that the proportion of South Asians dying from CVD is the highest of any other group in 1984-88. This indicated that South Asians are most likely to die from CVD than any other cause. The mortality data in our study did indicate that mortality occurs for South Asians proportionally more at a younger age than for all other Canadians. Possible reasons for this discrepancy could be due to the immigration process in which healthier people are more likely to be selected as immigrants and/or the better health services available in Canada may prolong the life of those diagnosed with CVD.

5. CONCLUSION

CVD mortality rates and trends differ for the various first generation ethnic groups in Canada. Approximately one in five Canadians is a first generation immigrant therefore it is important to consider ethnic background in the interpretation of health statistics and the planning of health services in Canada. Lifestyle plays an important role in determining the amount of exposure to the risk factor(s) and the resulting level of health. Most immigrants tend to carry over their cultural habits such as food choices and smoking behaviour for better or worse to their adopted country. Changes to these lifestyle patterns may be necessary in order to reduce the risk factors associated with CVD and other diseases.

In order to analyze health data multi-cultural and multi-ethnic society such as Canada, consideration should be given to collecting ethnic information in all information gathering related to health data and vital statistics. One of the major problems encountered in this study was the large gaps of data related to ethnic background. In particular, due to a lack of data, we were not able to investigate Canada's Native Indian population.

The investigation of differences in mortality rates between countries has traditionally aided in the identification of etiological factors. A recent publication (-----, 1989) has looked at the trends and determinates of CHD mortality internationally. Genetic, socioeconomic and lifestyle risk factors all influence international differences (Castelli, 1989). The genetic influences and variability for CVD risk factors are likely to be very complex and very difficult to unravel. There is a relationship with socioeconomic variables and CVD with more affluent countries having higher rates. However with the more affluent countries, low CVD rates are associated with higher social class. When one looks at changing social class a study by Marmot, 1989 found that in Britain moving to a higher social class is beneficial for some immigrant groups while also being detrimental in others.

With regard to risk factors it was shown that where there has been an increase in consumption of animal fat there has been an increase in CVD, where consumption did not change there was no change in CVD, and where consumption fell there was a decrease in CVD. This association was not confounded by smoking or alcohol use (Epstein, 1989). Campaigns to lower blood pressure in the United States have been mirrored by a 54% fall in stroke mortality (Blackburn, 1989). Changes in smoking habits do not help explain the difference in rates of decrease but in all likelihood have contributed to the decline. The situation is very complex and looking at changes in immigrant populations can play a role in elucidating it.

Table 1: Country Groupings

1. **NORTH AMERICA**
   Canada
   U.S.A.

2. **SOUTH AMERICA**
   Brazil
   Chile
   Mexico

3. **AFRICA**
   Continent

4. **AUSTRALIA***
   Continent

5. **ASIA:**
   6. CHINA
   7. JAPAN*
   8. SOUTH ASIA
      India
      Pakistan

9. **EUROPE:**

   10. SCANDINAVIA
   Denmark
   Finland
   Norway
   Sweden

   11. EASTERN EUROPE
   Czechoslovakia
   Hungary
   Poland
   Romania
   USSR
   Yugoslavia

   12. WESTERN EUROPE
   Austria
   France
   Greece
   Italy
   Spain
   United Kingdom
   Belgium
   Germany
   Ireland
   Portugal
   Switzerland

* Dropped because of low numbers.

Table 2: Data Availability

TIME PERIOD	NUMBER OF DEATHS	POPULATION FIGURES
1964-68	Available	Short census/missing place of birth
1969-73	Available	Available
1974-78	Missing: Ontario all yrs B.C. 3 yrs Manitoba 1 yr	Short census/missing place of birth
1979-83	Missing: Ontario 1 yr Manitoba 4 yrs Alberta 2 yrs	Available
1984-88**	Missing: Alberta 1 yr	Short census/place of birth available

* by place of birth, age, sex

** 1988 missing all figures, deaths & population

Table 3: Rank Ordering of ASMR - CVD

CANADIAN IMMIGRANTS	COUNTRY OF BIRTH
Scandinavia	Eastern Europe Scandinavia
North America	Western Europe
Eastern Europe	North America
Western Europe	
South Asia	South America
China	Hong Kong
South America	Sri Lanka

Table 4: Rank Ordering of ASMR - CVD

CANADIAN IMMIGRANTS	COUNTRY OF BIRTH
Scandinavia	Eastern Europe
Eastern Europe	Western Europe
Western Europe	Scandinavia
North America	Western Europe
South Asia	South America
China	Hong Kong
South America	Sri Lanka

REFERENCES

Balfe, D.L., W.J. Steinberg, H.G.V. Kustner. (1988), Comparison of the Decline in the Ischaemic Heart Disease Mortality Rate in the RSA with that in Other Western Countries. South African Medical Journal, 74, 551-553.

Blackburn, Henry. (1989), Trends and Determinants of CHD Mortality: Changes in Risk Factors and Their Effects. International Journal of Epidemiology, 18-S1, 210-215.

Canadian Censuses of Population (1971 and 1986). Statistics Canada.

Canadian Mortality Database, Marriage Statistics, Canadian Centre for Health Information. Statistics Canada.

Castelli, William. (1989), Determinants of CHD Mortality: Genetic, Socioeconomic, Lifestyle and Risk Factor Influences: An Overview. International Journal of Epidemiology, 18-S1, 180-182.

Department of National Health and Welfare. (1973), Nutrition Canada Survey (Cat. no. H58-36), Information Canada.

Employment and Immigration Canada. (1988), Quarterly Statistics-Immigration. Immigration Statistics, Employment and Immigration Canada.

Epstein, Frederick. (1989), The Relationship of Lifestyle to International Trends in CHD. International Journal of Epidemiology, 18-S1, 203-209.

----- (1989), Trends and Determinants of Coronary Heart Disease Mortality: International Comparisons. International Journal of Epidemiology, 18-S1.

Health and Welfare Canada. (1989), Canadian Blood Pressure Survey. (Cat. no. H39-143). Minister of Supply and Services, Ottawa.

Health and Welfare Canada. (1981), Canada Health Survey. The Health of Canadians: Report of the Canada Health Survey. Hull, Quebec. Minister of Supply and Services, Canada.

Hughes, L.O., V. Raval, E.B. Raftery. (1989), First Myocardial Infarctions in Asians and White Men. British Medical Journal, 298, 1345-50.

Hughes, L.O., J.H. Cruickshank, J. Wright, E.B. Raftery. (1989), Disturbances of Insulin in British Asian and White Men Surviving Myocardial Infarction. British Medical Journal, 299, 537-541.

Marmot, Michael. (1989), Socioeconomic Determinants of CHD Mortality. International Journal of Epidemiology, 18-S1, 196-202.

McKeigue, P.M., G.J. Miller, M.G. Marmot. (1989), Coronary Heart Disease in South Asians Overseas: A Review. Journal of Clinical Epidemiology, 42-7, 597-609.

Nair, C., H. Colburn, D. McLean, A. Petrasovits. (1989), Cardiovascular Disease in Canada. Health Reports, 1-1:1-22.

Salonen, Jukka T. (1989), Non-Insulin Dependent Diabetes and Ischemic Heart Disease. British Medical Journal, 298, 1050.

Statistics Canada. (1985), The Labour Force, June 1985 (Cat. no. 71-001). Minister of Supply and Services, Ottawa.

Thom, Thomas J. (1988), International Mortality from Heart Disease: Rates and Trends. International Journal of Epidemiology, 18-S1, 20-28.

World Health Organization. (1988), Noncommunicable Diseases: A Global Problem. World Health Statistics Quarterly Report, 41.

PART 6

DEMOGRAPHY

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

SOME STATISTICAL METHODS FOR PANEL LIFE HISTORY DATA

J.D. Kalbfleisch and J.F. Lawless[1]

1. INTRODUCTION

In fields such as demography, economics, medicine and sociology it is common to study multi-state life history processes of individuals from some population (e.g. Hoem 1985, Kalbfleisch and Lawless 1988, Tuma and Hannan 1984). In particular, suppose that, over time, N individuals move independently among k states $\{1,2, \ldots, k\}$. The states may, for example, represent separate disease conditions, occupational categories, marital status, or socio economic indicators. Let $X_l(t)$ be the state occupied by the l th individual at time t and let $\mathbf{Z}_l$ be a vector of covariates observed on the l th individual. Conditionally upon $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$, the processes $\{X_l(t): 0<t<\infty\}$, $l = 1, \ldots, N$ are assumed to be independent.

Sometimes complete individual life histories are observed over certain time intervals, and in Section 3 we review briefly this situation. The main focus of this paper, however, is to consider situations in which the l th individual is observed at a prespecified set of time points $t_{l0}, \ldots, t_{lm_l}$ and, at these times, the state $X_l(t_{lj})$ is observed. No information on the trajectory between successive observation times is available for use. Data of this type are commonly referred to as panel data. For ease of exposition, we assume that $t_{lj} = t_j$, $j = 0, \ldots, m$ where $m_l = m$, $l = 1, \ldots, N$ and $t_0 = 0$. The methods presented, however, generalize easily to the case in which the observation times differ from individual to individual. Particular interest centres on the modelling of the processes $X_l(t)$ so as to describe adequately the variation observed in the data. Comparison of processes for individuals in different groups and the assessment of covariates is often important.

In section 2, we review Markov and semi-Markov processes which are widely used as models for life history processes. In section 3, we review existing methods for situations where the processes $X_l(t)$ are observed continuously in time, and note some additional problems. In this case, there are several techniques available, based on a broad choice of models. Section 4 discusses methods for panel data at some length. Here convenient models and methods are much more limited. Our objective is to describe some feasible methods of analysis; in particular, we present procedures for time-homogeneous Markov models and simple extensions of them, including the incorporation of fixed covariates, unobservable heterogeneity, and time-dependence. We conclude in section 5 with some remarks on the usefulness of panel data and on related problems such as panel surveys.

2. SOME MODELS FOR LIFE HISTORY PROCESSES

Consider a continuous time process $\{X(t): 0<t<\infty\}$ with state space $S = \{1,2, \ldots, k\}$. For convenience, we suppose that the process enters state J_0 at time $T_0 = 0$. Let $M(t)$ be the number of transitions in $(0,t)$, T_r be the time of the r th transition, and J_r the state occupied immediately following the r th transition. Modelling can be accomplished through use of the instantaneous transition intensities,

$$\lambda_j(t; J_r, T_r, r = 0, \ldots, M(t)) = \lim_{\Delta t \to 0} P\{T_{M(t)+1} \epsilon (t, t+\Delta t), J_{M(t)+1} = j \mid J_r, T_r, r = 0, \ldots, M(t)\}/\Delta t.$$

Two special cases are of particular interest. Markov processes, for which

$$\lambda_j(t; J_r, T_r, r = 0, \ldots, M(t)) = q_{ij}(t) \tag{2.1}$$

where $J_{M(t)} = i \neq j$, specify that the transition intensities depend only on the calendar time, or the time elapsed since the time origin of the process. The special case of a homogeneous Markov process has

$$q_{ij}(t) = q_{ij} \tag{2.2}$$

independent of t and will prove a particularly useful model with panel data. It is convenient to let

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1.

$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$, $i = 1, \ldots, k$ and $Q(t) = (q_{ij}(t))_{k \times k}$ in the non homogeneous case or $Q = (q_{ij})_{k \times k}$ in the homogeneous case. For the homogeneous processes, it can be shown that

$$P(t) = e^{Qt} = I + Qt + Q^2t^2/2! + \cdots \tag{2.3}$$

where $P(t) = (p_{ij}(t))_{k \times k}$ and $p_{ij}(t) = P\{X(t)=j \mid X(0)=i\}$. See, for example, Cox and Miller (1965).

In the semi-Markov process, it is supposed that the transition intensities depend only on the time elapsed in the current state. Thus,

$$\lambda_j(t; J_r, T_r, r = 0, \ldots, M(t)) = \pi_{ij}(x) \tag{2.4}$$

where $J_{M(t)} = i$ and $x = t - T_{M(t)}$. For a review of such processes see, for example, Cinlar (1969).

More general models can also be considered. For example, the transition intensities may be functions of both the sojourn times x and the calendar time t. Such processes are called nonhomogeneous semi-Markov processes and provide a very flexible class of models. We shall later consider a mover-stayer model. In this case, it is assumed that each individual in state i at time 0 has a chance s_i of being a stayer and confined to state i for the full duration. With the complementary probability $1 - s_i$, $X(t)$ is a Markov process with intensities $q_{ij}(t)$.

There are extensions to regression models. If $\mathbf{Z}$ is a vector of covariates, it is natural to consider the Markov models with intensities

$$q_{ij}(t; \mathbf{Z}) = q_{ij}^0(t)\exp(\mathbf{Z}'\beta_{ij}) \qquad i \neq j \tag{2.5}$$

where $q_{ij}^0(t)$ is a baseline intensity function which applies when $\mathbf{Z} = 0$ and β_{ij} is a vector of regression parameters. (It is possible in (2.5) to replace $\exp(\mathbf{Z}'\beta_{ij})$ with an alternative relative risk function $r(\mathbf{Z}'\beta_{ij})$, but we shall use throughout the exponential relative risk function.) Semi-Markov regression models can be obtained in a similar way. More generally, and as discussed in Section 3, it is possible to allow the covariates to vary with time. Thus, for the Markov models, we may consider a covariate $\mathbf{Z}_t$ at time t that can depend on the process up to time t. Thus $\mathbf{Z}_t$ may include measured covariates, products between the covariates and time, information on previous states occupied, or the time $x = t - T_{M(t)}$ in the present state.

With panel data it is usually hard to deal with anything except Markov processes with fixed covariates $\mathbf{Z}$. With continuously observed life histories, however, a wide range of models are rather easily handled. We briefly review some methods in the next section.

3. METHODS FOR CONTINUOUS LIFE HISTORIES

If individuals are followed prospectively (forward over time) and their continuous life histories observed, then likelihood functions are readily obtained and methods of inference are straightforward. In particular, the Markov models in (2.5) can be generalized to multiplicative intensity models where the intensities are of the form

$$q_{ij}(t; \mathbf{Z}_t) = q_{ij}^0 \exp(\mathbf{Z}_t'\beta) \tag{3.1}$$

with $q_{ij}^0(t)$ a baseline intensity and $\mathbf{Z}_t$ a vector of possibly time-dependent covariates. For this model, analyses based on partial likelihood as described by Cox (1972, 1975) can be employed; Andersen and Borgan (1985) review this area for life history processes and Andersen (1985) discusses some economic applications. There seems to be considerable scope for the application of these methods in socio-economic studies, particularly when there are time-dependent covariates. For example, in employment-unemployment studies, transition intensities typically depend on fixed covariates associated with the individual, on calendar time t and covariates that vary with t (e.g. economic conditions), and on the life history of the individual such as the length of the sojourn in the state currently occupied. All of these are readily studied within (3.1). Our objective here is not to study methods for continuous observation, but we mention several areas where further development would be useful:

i) In many applications, individual life histories are partly determined retrospectively. For example, individuals may be sampled at some point in time and their recent life histories reconstructed, as might be done in determining the duration of the current unemployment spell or recent utilization of health care facilities. Aside from data accuracy problems, it is essential to account for selection bias by using appropriate likelihoods (e.g. see

Hoem, 1985, Kalbfleisch and Lawless, 1988). Problems of both design and analysis arise.

ii) Unobservable heterogeneity may be incorporated in models such as (3.1), for example by multiplying the transition intensities by random variables α_{ij}. Even for very simple models, however, this can be rather complicated; Andersen (1985) suggests one approach in conjunction with the analysis based on partial likelihood. Another problem is the impossibility of distinguishing between unobservable heterogeneity and certain types of time-dependence in a homogeneous population; further insights on this would be valuable.

iii) More exploration and application of models that incorporate a dependence both on calendar time and on duration of sojourn in the current state would be valuable. Experience in using such models with social and economic processes is at present limited.

4. PANEL DATA

4.1 Estimation for Homogeneous Markov Processes

The wide range of techniques available with continuous data contrasts with the limited methods available for the analysis of panel data. In this section, we review methods based on homogeneous Markov models and some simple extensions of them.

Let $\mathbf{Z}_l$ be a vector of covariates $(Z_{l1}=1, \ldots, Z_{lp})$ associated with the l th individual. The process $X_l(t)$ is taken to be a homogeneous Markov process with transition intensities

$$q_{ij}(\mathbf{Z}_l) = \exp(\mathbf{Z}_l' \beta_{ij}), \quad i \neq j \tag{4.1}$$

where $\beta_{ij}' = (\beta_{ij1}, \ldots, \beta_{ijp})$ is a vector of regression coefficients. Note that $\exp(\beta_{ij1})$ is the baseline transition intensity when $Z_{l2} = \cdots = Z_{lp} = 0$. We suppose that the vectors β_{ij} are functions of the parameter vector $\theta = (\theta_1, \ldots, \theta_q)$. In most applications, only a few of the components of β_{ij} are non zero and typically θ is of moderate dimension.

Let

$$P(t \mid \mathbf{Z};\theta) = \exp(Q(\mathbf{Z};\theta)t) = \sum_{j=0}^{\infty} \{Q(\mathbf{Z},\theta)t\}^j / j! \tag{4.2}$$

be the matrix of transition probabilities across an interval of length t. Finally, let $\mathcal{I}_{ijl}$ be the set of individuals observed to move from i to j in the interval $(t_{l-1}, t_l]$, $l = 1, \ldots, m$, $i, j = 1, \ldots, k$. The likelihood based on these data is

$$L(\theta) = \prod_{l=1}^{m} \prod_{i=1}^{k} \prod_{j=1}^{k} \prod_{r \in \mathcal{I}_{ijl}} p_{ij}(t_l - t_{l-1} \mid \mathbf{Z}_r; \theta). \tag{4.3}$$

In obtaining the maximum likelihood estimate of θ from (4.3), we follow the algorithm presented in Kalbfleisch and Lawless (1985). For a given value of $\mathbf{Z}$, and a specified value of θ, suppose that $Q(\mathbf{Z};\theta)$ has distinct eigenvalues $d_1 = 0$, $d_j(\mathbf{Z};\theta)$, $j = 2, \ldots, k$. Then

$$Q(\mathbf{Z};\theta) = A(\mathbf{Z};\theta)D(\mathbf{Z};\theta)A(\mathbf{Z};\theta)^{-1}$$

where the columns of A are the right eigenvalues of $Q(\mathbf{Z};\theta)$ and

$$P(t \mid \mathbf{Z};\theta) = A(\mathbf{Z},\theta)e^{D(\mathbf{Z},\theta)t} A(\mathbf{Z};\theta)^{-1}$$

where $D = diag(d_1, \ldots, d_k)$. This allows simple computation of the entries $p_{ij}(t \mid \mathbf{z},\theta)$ in the likelihood and the score vector $\partial \log L / \partial \theta$. The first derivatives in the score vector,

$$\frac{\partial}{\partial \theta} p_{ij}(t \mid \mathbf{Z};\theta) \tag{4.4}$$

can also be obtained using an algorithm due to Jennrich and Bright (1976). Bates and Watts (1988, Appendix A) discuss computational methods where the eigenvalues are not distinct. Kalbfleisch and Lawless (1985) show that a variation on the scoring algorithm for finding the mle's requires calculation of the first derivatives (4.4) only. At convergence a simple variance estimator for $\hat{\theta}$ is available.

It should be noted that these methods require separate calculations for each distinct covariate value. Thus, if the number of distinct covariate sets in the sample is large, these methods require a lot of computation. Panel data often do not support very detailed modelling of covariate effects, however, so in many situations the amount of computation is not excessive.

4.2 Incorporating Unobservable Heterogeneity

Even after modelling the dependence of transition intensities on explanatory variables, we may find that a homogeneous Markov model is inadequate. If Markov models for individuals are thought reasonable, one approach is to model "unobservable heterogeneity" in individual transition intensities through the use of random effects. We indicate how certain simple but useful models can be fitted.

A comprehensive Markov model for individual transition intensities would have transition intensity matrices $Q(\mathbf{Z}_l ;\theta \mid \alpha_l)$ where α_l is an unobservable vector of random effects associated with the l'th individual. In some situations special types of random effects may suggest themselves. We consider here only the simple though useful family of models for which

$$Q(\theta \mid \alpha_l) = \alpha_l Q^0(\theta), \tag{4.5}$$

where $Q^0(\theta)$ is a baseline intensity matrix and the α_l's are independent random variables with distribution function $G(\cdot)$. For simplicity of discussion, we suppose there are no fixed covariates $\mathbf{Z}_l$.

In order to write down the unconditional probability

$$Pr\{X_l(t_0), \ldots, X_l(t_m)\} = \int_0^\infty Pr\{X_l(t_1), \ldots, X_l(t_m) \mid X_l(t_0),\alpha\} Pr\{X_l(t_0) \mid \alpha\} dG(\alpha) \tag{4.6}$$

we need to specify the joint distribution of $X_l(t_0)$ and α_l. Defining $\eta_l = Pr\{X_l(t_0)=i\}$ and $G_i(\alpha)$ as the conditional distribution function of α_l given that $X_l(t_0) = i$, we can rewrite (4.6) as

$$\eta_i \int_0^\infty Pr\{X_l(t_1), \ldots, X_l(t_m) \mid X_l(t_0)=i,\alpha\} dG_i(\alpha) \tag{4.7}$$

when $X_l(t_0) = i$. In what follows we disregard the η_i's and estimate θ and the parameters of the $G_i(\alpha)$'s from the likelihood arising from the second term in (4.7). The model (4.5) implies that

$$P(t ;\theta \mid \alpha) = P^0(\alpha t ;\theta),$$

where $P^0(t ;\theta) = \exp\{Q^0(\theta)t\}$. The likelihood contribution from individual l is therefore

$$L_l = \int_0^\infty \{\prod_{r=1}^m p^0_{i_{r-1}i_r}(\alpha w_r)\} dG_i(\alpha), \tag{4.8}$$

where $i_{r-1} = x_l(t_{r-1})$, $i_r = x_l(t_r)$ and $w_r = t_r - t_{r-1}$ $(r = 1, \ldots, m)$. Note that it may not be clear how to model the $G_i(\alpha)$'s in many situations.

When the $G_i(\alpha)$'s are discrete the maximization of the likelihood arising from (4.8) can be handled with the algorithms discussed in section 4.1. As a simple but useful illustration we consider the so-called Mover-Stayer Model, which arises as the special case with $\alpha = 0, 1$ only. Blumen et al. (1955) introduced the discrete time version and Frydman (1984) discussed maximum likelihood estimation in the discrete model. We define $s_i = Pr\{\alpha_l=0 \mid X_l(t_0)=i\} = 1-Pr\{\alpha_l=1 \mid X_l(t_0)=i\}$ and consider joint estimation of $\mathbf{s} = (s_1, \ldots, s_k)$ and θ. The likelihood function is the product over $l = 1, \ldots, N$ of terms (4.8). Noting that $p_{ij}(0;\theta) = \delta_{ij} = I(i=j)$ and defining $n_{ijr} = \#\{l : X_l(r-1)=i, X_l(r)=j\}$, $n_i^* = \#\{l : X_l(r)=i$ for $r = 0,1, \ldots, m\}$ and $N_{ir} = \#\{l : X_l(r)=i\}$, we can rewrite the likelihood as

$$L(\theta, s) = \prod_{i,j=1}^{k} \prod_{r=1}^{m} p_{ijr}^{n_{ijr} - \delta_{ij} n_i^*} \prod_{i=1}^{k} [s_i + (1-s_i) \Pi_i]^{n_i^*} (1-s_i)^{N_{i0} - n_i^*} \tag{4.9}$$

where p_{ijr} stands for $p_{ij}(w_r; \theta)$ and $\Pi_i = \prod_{r=1}^{m} p_{ii}(w_r; \theta)$.

The first derivatives of $\log L$ may be shown to equal

$$\frac{\partial \log L}{\partial \theta} = \sum_{ijr} n_{ijr} p_{ijr}^{-1} p_{ijr}' - \sum_i \frac{n_i^* s_i}{s_i + (1-s_i)\Pi_i} (\partial \log \Pi_i / \partial \theta) \tag{4.10}$$

$$\frac{\partial \log L}{\partial s_i} = \frac{n_i^*(1-\Pi_i)}{s_i + (1-s_i)\Pi_i} - \frac{N_{i0} - n_i^*}{1-s_i}, \tag{4.11}$$

where $p_{ijr}' = \partial p_{ijr} / \partial \theta$ and we note that $\partial \log \Pi_i / \partial \theta = \sum_{r=1}^{m} p_{iir}^{-1} p_{iir}'$. All quantities in (4.9) and (4.10) can be obtained with the algorithms in Kalbfleisch and Lawless (1985) that were mentioned in Section 4.1.

By direct calculation, the Fisher information matrix has entries

$$E\left(-\frac{\partial^2 \log L}{\partial s_i^2}\right) = \frac{N_{i0}(1-\Pi_i)}{(1-s_i)(s_i + (1-s_i)\Pi_i)}$$

$$E\left(-\frac{\partial^2 \log L}{\partial s_i \partial \theta}\right) = \frac{N_{i0}\Pi_i}{s_i + (1-s_i)\Pi_i} (\partial \log \Pi_i / \partial \theta)$$

$$E\left(-\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right) = \sum_{i,j,l} (E(N_{i,l-1}) - N_{i0} s_i) p_{ijl}^{-1} p_{ijl}' (p_{ijl}')'$$

$$- \sum_i \frac{N_{i0} s_i (1-s_i)\Pi_i}{s_i + (1-s_i)\Pi_i} (\partial \log \Pi_i / \partial \theta)(\partial \log \Pi_i / \partial \theta)'.$$

An estimate of this is simply obtained by substituting $n_{i,l-1}$ for $E(N_{i,l-1})$.

As for the homogeneous case in section 4.1, Fisher's scoring algorithm provides a simple method for fitting the data, and only first derivatives of the $p_{ij}(w_r; \theta)$'s with respect to θ are needed. Note that $\partial \log L / \partial s_i = 0$ yields the s_i's in terms of θ, so that some simplification is possible.

More general mover-stayer models can also be fitted which allow, for example, dependencies of s_i on covariates, or regression models for the transition intensities q_{ij}.

Models of the form (4.5) provide a useful extension to homogeneous Markov models even though they may represent a considerable oversimplification. More detailed random effects modelling, moreover, usually leads to such difficult estimation problems with panel data that even considering it seems questionable. For example, consider a two-state model with random effects α_1 and α_2 so that, conditional on α_1 and α_2, an individual has transition intensities $\alpha_1 q_{12}^0$ and $\alpha_2 q_{21}^0$. Even if α_1 and α_2 are assumed independent, estimation is forbidding. More realistically, α_1 and α_2 would not usually be independent; estimation in this case would be complicated even for continuously observed life histories.

4.3 Incorporating Non Homogeneous Behaviour

i) The methods can be extended to allow for some types of non homogeneous Markov models. If, for example

$$Q(t \mid \mathbf{Z}; \theta) = Q(\mathbf{Z}; \theta) h(t), \tag{4.12}$$

then use of the operational time scale $s = \int_0^t h(u)du$ gives rise to a homogeneous process for $Y_l(s) = X_l(t)$, and the same calculations can then be applied for known $h(t)$. If $h(t) = h(t; \lambda)$ depends on a vector of parameters λ, then λ can be estimated by exploring the profile likelihood for λ obtained by maximizing over the regression parameters θ for each given λ. An open question of some interest relates to the case where the model (4.12) is assumed, but $h(t)$ is arbitrary. It seems likely that methods can be developed to handle this case.

ii) An alternative approach to incorporate non homogeneity is to allow the Markov matrix Q to change at prespecified time points. Thus it is assumed that

$$Q(t \mid \mathbf{Z};\theta) = Q_r(\mathbf{Z};\theta) \qquad a_{r-1} \leq t < a_r$$

where $r = 1, \ldots, s$, $a_0 = 0$ and $a_s = \infty$. Calculations are simplest if change points are assumed to occur at some of the observation times $t_1, \ldots, t_m$, but this model can be fitted quite generally. This allows incorporation of a time varying baseline intensity matrix; a model of the form (2.5) can be fitted.

iii) Tests for time homogeneity could be based on either of the nonhomogeneous models discussed above. A further possibility follows a suggestion of de Stavola (1988) who considers a special case of the following. Suppose that there are no covariates and consider a model of the form

$$Q(t) = Q + H\gamma t$$

where $Q = (q_{ij}(\theta))$ as before, and H is a given matrix which specifies components of $Q(t)$ which may be nonhomogeneous. For example, if transition rates from r to s were suspected to vary with time, one might consider $H = (h_{ij})_{k \times k}$ where $h_{rs} = 1$, $h_{rr} = -1$ and $h_{ij} = 0$ otherwise. A score test of $\gamma = 0$ can be used to provide an assessment of the time homogeneity assumption versus alternatives in which the rate increases or decreases with time.

If n_{ijl} is the number of observed transitions from i to j on the interval t_{l-1}, t_l, the log likelihood function is

$$\log L = \sum_{i,j,l} n_{ijl} \log p_{ij}(t_{l-1}, t_l). \tag{4.13}$$

The transition probabilities can be written as product integrals,

$$P(t_{l-1}, t_l) = \prod_{u \in (t_{l-1}, t_l)} \{I + Qdu + H\gamma u du\} \tag{4.14}$$

$$= \lim \prod_{l=1}^{M} \{I + [Q + \gamma H u_i]\Delta u_i\}$$

where $u_0 = t_{l-1} < u_1 < \cdots < u_M = t_l$, $\Delta u_i = u_i - u_{i-1}$ and the limit is taken as $M \to \infty$, $\max \Delta u_i \to 0$. To obtain a score statistic for testing $\gamma = 0$, we require evaluation of the derivative of (4.14) with respect to γ. Evidently, this derivative can be written as

$$\left. \frac{\partial p_{ij}(t_{l-1}, t_l)}{\partial \gamma} \right|_{\gamma=0,\hat{\theta}} = A \int_{t_{l-1}}^{t_l} e^{D(s-t_{l-1})} A^{-1} H A e^{D(t_l - s)} ds A^{-1} \tag{4.15}$$

where $D = D(\hat{\theta})$ is the diagonal matrix of eigenvalues at the m.l.e. under the homogeneous model $(\gamma=0)$ and $Q(\hat{\theta}) = ADA^{-1}$ as before. The expression (4.15) can be further simplified by noting that the integrand is a matrix of exponential functions, and can be simply evaluated.

Estimation of the variance of the score statistic requires evaluation of the Hessian matrix corresponding to (4.15) at $\gamma = 0$ and $\hat{\theta}$. Calculations here are more complicated but proceed in a manner similar to those above. De Stavola (1988) gives an example of this approach in a simple case where the $p_{ij}(s, t)$'s can be obtained algebraically.

5. SOME OTHER POINTS

Panel studies are mainly useful for studying individual life histories when the life histories can be modelled as Markov processes. If not, and in particular if transition intensities depend strongly on time spent in a state, panel studies are not particularly helpful unless observation times are close together. We remark that in some types of studies it may be possible to retrospectively ascertain, or at least estimate, the life history of individual subjects between observation times. The effect of possibly inaccurate information then has to be weighed against the benefits of more complete life histories.

There are several aspects of panel data analysis that have not been addressed. One concerns panel surveys where at each observation time t_i $(i = 0, 1, \ldots, m)$ some new individuals are introduced into the study and some individuals leave it. For example, labour force surveys often employ so-called rotating panels (e.g. see Trivelatto and Torelli 1989). A main objective of such studies is often to estimate the proportion of a population in the various states at $t_0, t_1, \ldots, t_m$, but information about transitions and sojourn times in states is also available. Many analyses of such studies rely on sample survey methods (e.g. Smith and Holt 1989); it would be interesting to see more life history model-based analyses carried out. Interesting design issues also arise, for example concerning the use of cross-sectional vs. longitudinal information (cf. Lawless and McLeish 1984).

Panel studies typically involve the observation of a fairly large number of individuals at relatively few time points $t_0, t_1, \ldots, t_m$. When processes are in equilibrium there may be substantial information contained in the initial distribution of individuals' states at t_0. For Markov processes this is readily incorporated: we sketch the approach for homogeneous models.

Let $Q(\theta)$ be the $k \times k$ transition intensity matrix and $\pi = \pi(\theta)$ the corresponding $k \times 1$ equilibrium distribution; π is the unique solution to $Q'\pi = 0$, $\pi_1 + \cdots + \pi_k = 1$. The likelihood function based on the distributions $X_l(t_0), \ldots, X_l(t_m)$ for individuals $l = 1, \ldots, n$ is

$$L(\theta) = \prod_{i=1}^{k} \pi_i(\theta)^{n_i(0)} \prod_{ijr} P_{ij}(w_r; \theta)^{n_{ijr}}, \tag{5.1}$$

using the same notation as in Section 4.1. To maximize (5.1) by solving the likelihood equations we need to compute $\partial \pi / \partial \theta'$. Let Q_1 be the $k \times (k-1)$ matrix obtained from the first $k-1$ columns of Q and let $\pi_1 = (\pi_1, \ldots, \pi_{k-1})'$. Then Q_1' may be considered as a function $F(\theta, \pi_1)$ defining π_1 implicitly in terms of θ via $F(\theta, \pi_1) = 0$, as in Kalbfleisch and Lawless (1985, Appendix B). Implicit differentiation of $F(\theta, \pi_1)$ with respect to θ shows that

$$\partial \pi_1 / \partial \theta' = -B(\theta)^{-1} C(\theta),$$

where $B(\theta)$ is $(k-1) \times (k-1)$ with (i, j) element $q_{ji}(\theta) - q_{ki}(\theta)$ and $C(\theta)$ is $(k-1) \times p$ with j'th column $(\partial Q_1' / \partial \theta_j)\pi_1$. It may be checked that the scoring method for solving $\partial \log L / \partial \theta = 0$ requires only first derivatives of Q and π_1 with respect to θ, and hence the algorithms described in Section 4.1 allow the calculation of all necessary quantities.

Finally, we remark that if individuals are observed at equally spaced time points $t_0, t_0 + \tau, t_0 + 2\tau, \ldots$. simpler Markov chain models are often employed (eg. Bishop et al. 1975, ch. 7). For a homogeneous model, this allows the estimation of transition probabilities $p_{ij}(\tau)$ but does not in general yield estimates of more general transition probabilities, intensities or sojourn distributions (cf. Kalbfleisch and Lawless 1985, Section 7). This may not be a drawback for certain applications.

REFERENCES

Andersen, P.K. and Borgan, O. (1985). Counting Process Models for Life History Data: A Review (with discussion). *Scandinavian Journal of Statistics 12*, pp. 97-158.

Andersen, P.K. (1985). Statistical Models for Longitudinal Labor Market Data Based on Counting Processes. Chapter 6 in Heckman, J.J. and Singer, B. (eds.). *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.

Bates, D.M. and Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.

Blumen, J., Kogan, M. and McCarthy, P.J. (1955). *The Industrial Mobility of Labor as a Probability Process*. Cornell Studies of Industrial and Labor Relations, Vol. 6. Ithaca, N.Y.: Cornell University Press.

Cinlar, E. (1969). Markov Renewal Theory. *Advances in Applied Probability 1*, 123-187.

Cox, D.R. (1972). Regression Models and Life Tables (With Discussion). *Journal of the Royal Statistical Society (B), 34*, 187-220.

Cox, D.R. (1975). Partial Likelihood. *Biometrika, 62*, 269-276.

Cox, D.R. and Miller, H.D. (1985). *The Theory of Stochastic Processes*. London: Methuen (ch. 4).

de Stavola, B.L. (1988). Testing Departures From Time Homogeneity in Multistate Markov Processes. *Applied Statistics, 37*, 242-250.

Frydman, Halina (1984). Maximum Likelihood Estimation in the Mover-Stayer Model. *Journal of the American Statistical Association 79*, 632-639.

Hoem, J.M. (1985). Weighting, Misclassification, and Other Issues in the Analysis of Survey Samples of Life Histories, Chapter 5 in Heckman, J.J. and Singer, B. (eds.). *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge.

Jennrich, Robert I. and Bright, Peter B. (1976). Fitting Systems of Linear Differential Equations Using Computer Generated Exact Derivatives. *Technometrics, 18*, 385-392.

Kalbfleisch, J.D. and Lawless, J.F. (1985). The Analysis of Panel Data Under a Markov Assumption. *Journal of the American Statistical Association, 80*, 863-871.

Kalbfleisch, J.D. and Lawless, J.F. (1988). Likelihood Analysis of Multi State Models for Disease Incidence and Mortality. *Statistics in Medicine, 7*, 149-160.

Lawless, J.F. and McLeish, D.L. (1984). The Information in Aggregate Data from Markov Chains. *Biometrika, 71*, 419-430.

Smith, T.F.M. and Holt, D. (1989). Some Inferential Problems in the Analysis of Surveys Over Time. *Bulletin of the International Statistical Institute*, Vol. 53, 405-424.

Trivellato, U. and Torelli, N. (1989). Analysis of Labor Force Dynamics from Rotating Panel Survey Data. *Bulletin of the International Statistical Institute*, Vol. 53, 425-444.

Tuma, N.B. and Hannan, M.T. (1984). *Social Dynamics; Models and Methods*. New York: Academic Press.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

EVENT HISTORY ANALYSIS OF MARRIAGE AND DIVORCE IN CANADA

Geoff Rowe*

ABSTRACT

The 1984 Family History Survey provides retrospective data on the timing of births, marriage, divorce and work interruptions from a probability sample of Canadian males and females. The data comprises dates of events occurring in the time interval 1932-84. Individual event histories reconstructed from the dates represent about 4 million person months of experience.

Marital event histories are analyzed employing hazard regression. Regression results suggest that historical trends in marriage and divorce rates are associated with trends in male and female work patterns, and that trends in marital fertility have influenced trends in incidence of divorce.

The work reported in this paper was undertaken as a component of the development of a demographic microsimulation model (DEMOGEN) at Statistics Canada.

KEYWORDS: Marriage; Divorce; Event History

1. INTRODUCTION

1.1 Analyzing Marital Status Transitions

Marriage is an event most people eventually experience (at least once), yet there is considerable variation in its timing (e.g., age at marriage). Divorce is the final outcome in an increasing proportion of marriages, but patterns of marriage breakdown (i.e., separation or divorce) by marriage duration are not well understood (Burch & Madan, 1986). This paper presents an analysis of marital event histories indicating the close relations existing between the timing of marital status transitions and the pace of change in other areas of individuals' lives.

Descriptions of the timing of marital events can be given in terms of probabilities that an individual may change marital status within short time intervals. Such probabilities will be relatively large, if the expected duration until the occurrence of an event is relatively short (and correspondingly, the probabilities will be small, if expected durations are long). There is an advantage from studying probabilities rather than durations, since different types of durations can be involved in a single type of event. For example, divorce probabilities may reflect effects of both the duration of the marriage and the ages of children (i.e., birth intervals).

The structure of marital status transition probabilities is represented in the transition matrix. Within the matrix, certain combinations of initial statuses and outcomes will correspond to observable events, while other combinations may be deemed impossible. In this study, the transition matrix has the form given in Figure 1 below, for marital statuses defined as:

(1) SNGL - Never married and not in a common-law union
(2) CLU - Never married and in a common-law union
(3) MAR - Legal 1st marriage
(4) SEP - Permanent separation from the 1st marriage, but not legally divorced [1]
(5) DIV - Divorce or annulment of 1st marriage,

* Geoff Rowe, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario K1A 0T6

with X's denoting observable events, and -'s denoting impossible events.

FIGURE 1

Transition Matrix

INITIAL \ STATUS	OUTCOME STATUS				
	SNGL	CLU	MAR	SEP	DIV
SNGL	X	X	X	-	-
CLU	X	X	X	-	-
MAR	-	-	X	X	-
SEP	-	-	-	X	X
DIV	-	-	-	-	X

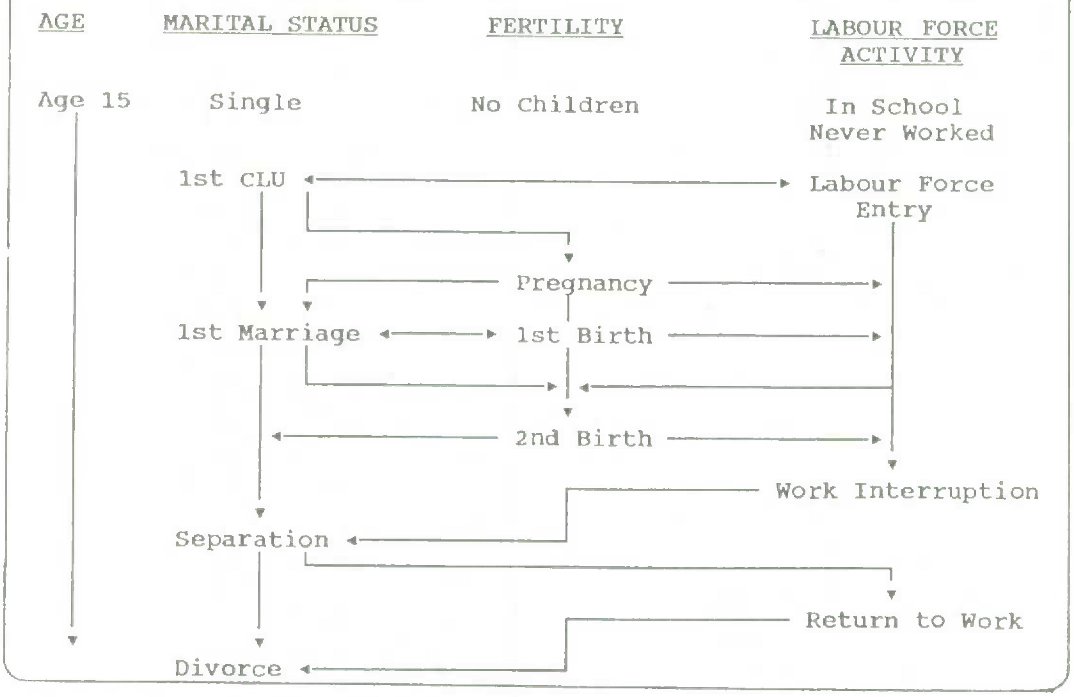
A priori, marital status transition probabilities would only be expected to remain constant with the passage of time in exceptional circumstances. Change in the probabilities would normally be expected in association with age or maturity. Change may also come about because of altered circumstances. For example, marriage may be more likely after labour force entry than before, regardless of age. In a study of conjugal union dissolution among Swedish women, Hoem (1989, p.2) emphasizes the relations among life events:

"As a woman's life unfolds, she continuously updates her life strategy under the influence of a highly dynamic system of resources, experiences, choices, restrictions, and chance outcomes. Her strategy determines her behaviour and produces what appears in the probability scheme as mutual causation between her educational and employment careers and her family history".

The following path diagram (Figure 2) illustrates some possible orderings of events (e.g., 1st conception prior to marriage) and possible paths of influence (e.g., return to work as a consequence of separation). This particular diagram is not intended to be typical in any sense. There are many alternative orderings and paths of influence, each of which could reflect a different life strategy.

FIGURE 2

Hypothetical Path Diagram of Concurrent Event Histories



The decade of life in which the first marriage usually occurs is generally also the decade in which schooling is completed and labour force entry first takes place, often in conjunction with separation from the parental home.

Each of these is a major step in the life course. Similarly, the terms or conditions of a marriage, once established, may change with the arrival of children or with the accumulating labour force experiences of marriage partners. For example, episodes of unemployment may have a transient or a cumulative effect which increases the risk of marriage breakdown. It thus seems essential that analysis of marital event histories give careful consideration of the potential for time-varying influences of concurrent histories of related events in individuals' lives.

In order to represent the dynamics involved in marital status transitions, each of the cells of the transition matrix will correspond to a separate regression equation which simultaneously accounts for aging and the effects of changing circumstances. Separate, but not necessarily independent, equations are called for in dealing with transitions involving competing risks (e.g., SNGL to MAR or CLU).

The path diagram also highlights the need for detailed data on the timing of events. For example, in order to resolve the question of whether CLU formation influences labour force entry or vice versa, the order of these events has to be determined. This will frequently require dates of events on a monthly rather than an annual basis. Such data can only be provided by special purpose, in-depth surveys such as the 1984 Family History Survey. The scarcity of such data has been cited as the principal reason for our "rudimentary" understanding of marriage and divorce in Canada (Balakrishnan, et.al., 1987).

1.2 Previous Findings

In many studies of marriage and marriage breakdown, the covariates examined have either been fixed (i.e., time invariant) individual characteristics (e.g., race or ethnicity) or have been treated as fixed (e.g., urban/rural residence or education level). This limitation may seriously reduce the value of some studies, in view of the dynamics illustrated in Figure 2.

In a study of recent Canadian census data, Genier et. al. (1987) found that mother tongue, region of birth and of residence, religion and education had significant effects on the mean age at marriage. Moreover, the effects differed between the 1971 and 1981 censuses, as well as among birth cohorts.

Parental socioeconomic status has had considerable attention as a possible predictor of the timing of marriage (e.g., Hogan, 1978) with generally inconclusive findings. However, education level is strongly related to marriage timing[2] with higher levels of educational attainment resulting in delayed marriage. Both labour force participation and income have been considered as factors influencing marriage decisions, although in differing ways for men and women (Teachman, et.al., 1987). Findings to date suggest that success in the labour force may accelerate marriage for men, but delay marriage for women.

Statistics on divorce typically gain more attention than do those on marriage. Undoubtedly, this is because divorce is still an unusual event (i.e., as yet, even the most pessimistic estimates indicate that most marriages do not end in divorce), while marriage is nearly universal (i.e., 90-95% eventually marry). Furthermore, the acceleration of divorce rates in Canada after 1968 (divorce legislation reform) may reflect an important social change.

Balakrishnan et.al. (1987) make use of data from the Canadian Fertility Survey to identify significant effects on risk of breakdown associated with age at marriage, marriage cohort, cohabitation before marriage, pre-marital fertility status, church attendance and place of residence (i.e., urban size class). Studies of American data have reported effects associated with the presence of children, and with employment or income (Teachman, et.al., 1987). The latter studies frequently indicated conflicting conclusions. However, Hannan and Tuma (1978) examined the impact of increased family income and were able to distinguish between effects of increasing a family's economic well being (i.e., reduced risk of breakdown) and effects of reducing the economic dependence of one partner on the other (i.e., increased risk of breakdown).

2. FAMILY HISTORY SURVEY DATA

The 1984 Family History Survey (FHS) provides data on concurrent marital, fertility and work histories for the first time from a probability sample of Canadians.

The FHS was conducted by Statistics Canada as a supplement to the Labour Force Survey. About 14,000 Canadian men and women (aged 18-64) were interviewed individually and were asked to give a detailed history of their marriages, common-law unions, separations and divorces (dates of events in year and month). In addition, dates were collected for entries and exits from care of children (natural, step or adopted), for the respondent's 1st job (full or part-time, but lasting 6 months or more and excluding work while a student) and for respondent's experience of prolonged periods (one year or more) of work interruption. The questions on work were given different treatment than others, in that the questions dealt with years of work and interruption durations in years (i.e. rather than the year and month of events). Further details and a general overview of the survey content are available in Burch (1985).

FHS data are available in the form of the dates reported by each respondent for 0-3 marriages, 0-6 CLUs, 0-15 children and 0-4 work interruptions. For the most part, these data were sufficient to determine the state of the respondent in respect of their marital, childcare and work status on a monthly basis. In preparation for analysis, the respondent file was converted to a monthly file on which each person month (after the 15th birthday) represented a separate record. This conversion produced an expansion of the FHS file from 14,004 responses to over 3.9 million person month records.

Tabulations of unweighted counts of marital status transitions, respondents and person years at risk are presented in Table 1. These and all subsequent results reported in this paper take no account of survey weights. Hoem (1985, p.258) notes:

"If [the survey design] is noninformative, then one may disregard the sampling plan and treat the sample of life histories as so many independent sample paths of stochastic processes with the probabilistic properties they would have had without the interference of survey sampling."

Where the weights have been used to estimate known population totals from FHS data (e.g., divorces in Canada by year of occurrence, Burch (1985)), the estimates have been poor. Hence, unweighted analysis was attempted first, and weighted regressions have not revealed large differences from unweighted.

TABLE 1
UNWEIGHTED COUNTS

Population At Risk	Respondents	Person Years At Risk [3] (Months/12)	Transition Outcomes	
SNGL			CLU	MAR
Never Worked	6545	22863.0	277	656
Ever Worked	6795	31167.0	1175	3311
CLU			SNGL	MAR
Never Worked	too few cases		-	-
Ever Worked	1202	3013.0	303	562
MAR			SEP	
	10456	181560.6	1389	
SEP			DIV	
	1248	4080.4	879	

Note that the transition structure indicated in Figure 1 has been further refined to distinguish between periods before or after labour force entry for each respondent.

3. PROPORTIONAL HAZARDS REGRESSION

Hazards models represent transition probabilities in continuous time (Cox and Oakes, 1984). The density and cumulative distribution functions (conditional on covariate vector X) of duration variable T are $f(T|X)$ and $F(T|X)$, respectively. The conditional transition probability (hazard) is defined as:

$$h(T|X) \equiv f(T|X)/(1-F(T|X)) = -d \log(1-F(T|X))/dT$$

(i.e., the limit of an occurrence/exposure ratio as the exposure duration approaches 0). From this definition, it follows that:

$$F(T|X) = 1 - \exp(- \int_0^T h(t|X)\, dt) = 1 - \exp(-H(T|X)),$$

where $H(T|X)$ is the cumulative hazard and consequently:

$$f(T|X) = h(T|X) \cdot \exp(-H(T|X)).$$

Thus, hazard models are equivalent to the specification of a density function for the duration variable.

Let L be the likelihood of a censored sample of duration observations with time-varying covariates $(X(t))$ and regression coefficients (β). The contribution to likelihood L from individual i given β is:

$$L_i = f(T_i|X_i(T),\beta)^{1-c_i} \cdot (1-F(T_i|X_i(T),\beta))^{c_i}$$

where T_i is the event date or censorship date for individual i. The value c_i=1 denotes censorship, while c_i=0 indicates an observed event. Maximizing $\log(L)= \Sigma_i \log(L_i)$ (i.e., for the sample of n individuals) is equivalent to maximizing hazards associated with observed events and minimizing cumulative hazards for censored observations. After rearrangement, the log likelihood is:

$$\log(L_i) = (1-c_i)\log(h(T_i|X_i(T),\beta)) - \int_0^{T_i} h(t|X_i(t),\beta)\, dt .$$

Maximization of L can be approximated by Poisson regression. The interval $(0,T_i)$ may be partitioned into small, non-overlapping intervals $(t_{i,j-1},t_{i,j}]$ (for this study, the intervals are months). Then the pseudo-observations $e_{i,j}$ represent whether or not an event occurred in each time interval j for each individual i:

$$e_{i,j} = \begin{cases} 0 , \text{ if } t_{i,j} < T_i \\ 0 , \text{ if } t_{i,j-1} < T_i \le t_{i,j} \;\&\; c_i=1 \\ 1 , \text{ if } t_{i,j-1} < T_i \le t_{i,j} \;\&\; c_i=0. \end{cases}$$

For purposes of estimation, the regression equation relates time-varying independent variables $(X_{i,j})$ to the expectations of the pseudo-observations:

$$E(e_{i,j}) = h_{i,j} = \exp(X_{i,j}\, \beta),$$

then $$\text{Log}(L) \approx \Sigma_{i,j} ([e_{i,j} \cdot \log(h_{i,j})] - h_{i,j}),$$

where $h_{i,j}$ is a piece-wise constant approximation to the hazard function (i.e., which is understood to be conditional on covariates and on β). The approximate likelihood has the same form as a Poisson likelihood (i.e., for pseudo-observations $e_{i,j}$), and may be maximized by iteratively reweighted least squares (McCullagh and Nelder, 1983) employing standard regression software.

4. SELECTED REGRESSION RESULTS

A variety of regression specifications have been attempted for each of the types of events identified in Figure 1. Given the total of 34 equations estimated, full detail cannot be provided here (but is available on request). The results reported in this section are selected aspects of the best fitting equations, and focus on two topics: (i) the effect of labour force entry on marriage and CLU formation, and (ii) the effects of children on risks of separation. The best fitting equations were selected on the basis of maximized likelihood, and from examination of generalized residuals (Cox and Oakes, 1984, pp.88-89).

An indication of the content of the best fitting regressions is given in Table 2, which provides estimated annual transition rates for selected combinations of independent variables. This table gives baseline rates which aid interpretation of subsequent tables of relative risks (RR) (i.e., hazard ratios).

TABLE 2
Annual Transition Rates/1000 From Regression Equations

INITIAL STATUS		EVENT	
SNGL			
Never Worked			
Aged 20, No Pregnancy		Marriage	CLU Formation
No Previous CLU	Male	1.8	1.1
	Female	4.7	2.0
Labour Force Entrant			
Aged 20, No Pregnancy			
No Previous CLU	Male	14.8	22.3
	Female	17.4	42.6
CLU			
Labour Force Entrant			
1st Year of the CLU			
Aged 20, No Pregnancy		Marriage	CLU Dissolution
No Previous CLU	Male	19.7	19.8
	Female	23.2	15.7
MAR			
After Divorce Legislation Reform			
Childless, Married at Age 21+			
No Preceding CLU			
1st Year of Marriage		Separation	
No Work Interruptions	Male	13.7	
No Work Since Marriage	Female	6.5	
SEP			
2nd Year of Separation		Divorce	
Preschool Child at Separation			
No New CLU		272.2	

The baseline rates in Table 1 compare reasonably well with appropriate age specific estimates based on vital statistics, however no vital statistics estimates with comparable definitions are possible.

4.1 Marriage and CLU Formation at Labour Force Entry

Analysis of the FHS data seems to indicate an independence effect, whereby labour force entry (and associated financial independence) marks the maximum marriage rate that a labour force cohort will typically experience. The only serious qualification being that the occurrence of a pregnancy will overwhelm all other effects for its duration. It is doubtful that the independence effect could entirely represent a causal relation. It may frequently happen that labour force entry is motivated by a decision to marry.

Evidence of the independence effect may be found in the annual transition rates from SNGL status (Table 2) which show a marked difference between those who have never entered the labour force and those who entered the labour force in the year they turned 20. The regression equations indicate the same kind of result for ages at labour force entry ranging from 15 to 35. Figure 3 compares hazards at labour force entry with hazards prior to labour force entry. The vertical axis is the relative risk computed from the hazard for labour force entrants divided by the hazard associated with the same age in the Never Worked status. Clearly, for ages above and below 20, the independence effect may be even greater than indicated is in Table 2.

Table 3 presents relative risks of union formation, either marriage or CLU formation, in years subsequent to labour force entry for SNGL persons. For each age at entry, the chance of marriage or CLU formation declines steadily thereafter. All else equal, marriage rates peak in the years immediately

FIGURE 3

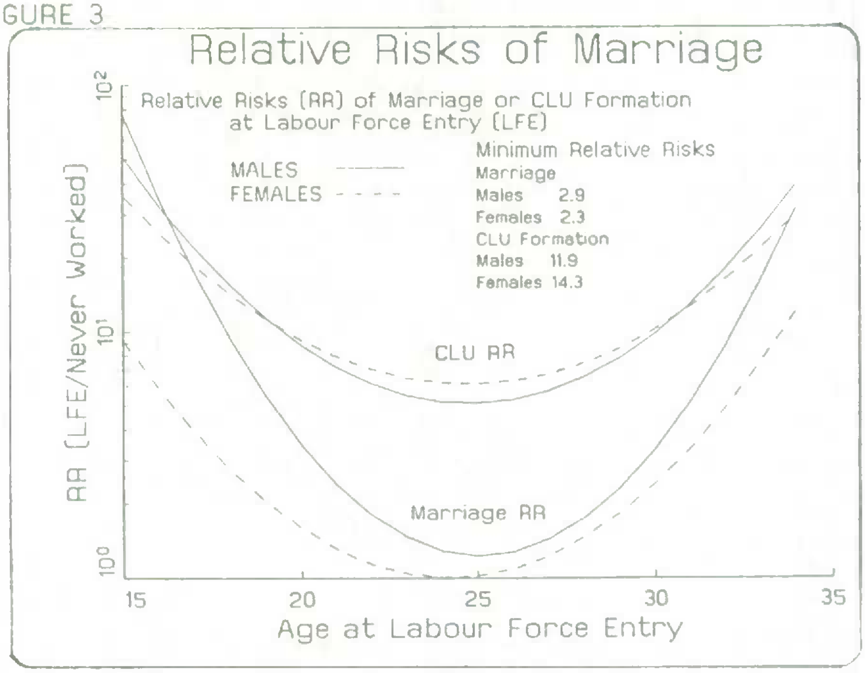


TABLE 3
Relative Risks of Marriage and CLU Formation by Work Status

Work History

Age at Labour Force Entry	Completed Years of Work							
	0	5	10	15	0	5	10	15
Males	Marriage				CLU Formation			
15	0.70	0.71	0.71	0.72	0.91	0.75	0.61	0.50
20	1	0.76	0.58	0.44	1	0.64	0.41	0.27
25	1.42	0.82	0.48	0.28	1.10	0.56	0.28	0.14
30	2.03	0.89	0.39	0.17	1.22	0.48	0.19	0.07
Females								
15	0.90	0.77	0.66	0.57	1.15	0.84	0.62	0.45
20	1	0.60	0.36	0.22	1	0.50	0.24	0.12
25	1.11	0.47	0.20	0.09	0.87	0.29	0.10	0.03
30	1.23	0.37	0.11	0.03	0.75	0.17	0.04	0.01

Concurrent Work Status

	Marriage	CLU Formation	Marriage	CLU Formation
	Males		Females	
Working	1	1	1	1
Not Working	0.26	0.71	0.80	0.70
Starting Work	0.45	0.56	0.51	0.61
Stopping Work	0.57	0.60	2.22	1.60
Returned to School	0.47	1.12	1.06	0.81

following labour force entry. However, the FHS work histories were given in years, rather than in years and months. The effect of rounding to the nearest year is that there are months with indeterminant work status (labelled Starting Work and Stopping Work) where the order of events (e.g., entry prior to marriage) can not be established precisely. Thus, it can not be determined whether the independence effect is associated with a brief period immediately following labour force entry, rather than a period spanning the year or two following entry.

Analysis of data for the stage following labour force entry resulted in a best fitting equation including age at entry and years of work as distinct variables with coefficients of differing sign (except female CLU formation). It is an important point that these variables have different effects, since their sum corresponds to current age in most cases (i.e., except where there have been spells of work interruption). This serves to indicate that marriage rates disaggregated by age group alone (i.e., birth cohorts) will conceal important heterogeneity among labour force cohorts.

Education histories were not collected in the FHS, however the question on the timing of the first job stipulated that work while a fulltime student should not be counted. Consequently, age at first job may be treated as a proxy for completed years of education. Thus, the data on age at first job and subsequent years of work are proxies for education and work experience, (human capital variables often employed by economists in analyzing wage rate differentials).

The observation that marriage chances typically increase with age at entry, but decrease with years of work suggests that earnings may not be associated with the timing of marriage decisions in a simple way (if at all); since education and work experience should both be positively associated with earnings. Nevertheless, the relative risk associated with the concurrent status Not Working (Table 3) indicates that work interruption results in a reduction in marriage chances (a more marked reduction for males than for females). In this respect, economic factors may have a direct role. Note that Not Working status is distinguished from the indeterminate statuses (Starting and Stopping Work) and does not include episodes where the reason given for the work interruption was a return to school.

4.2 Separation, Work and the Ages of Children

The "seven year itch" is part of North American folklore, as is the notion that children (especially young children) are the sole reason that some marriages hold together. The former has some empirical basis in the observation that divorce rates by duration of marriage often increase over the initial years of marriage and then decrease at longer durations. The implication seems to be that marriages may be viewed as, in some sense, "wearing out".

Another explanation of the duration pattern of divorce supposes that marriages may be classified as either stable or unstable, at the outset. Stable marriages have fixed, low risks of divorce. Unstable marriages have risks that steadily increase with time (i.e., following the honey moon period). Since, we can not normally distinguish between the types, we observe a divorce rate which is an average of two rates, but is initially dominated by unstable marriages (i.e., the average divorce rate increases). As unstable marriages are effectively "weeded out", the average divorce rate comes to be dominated by the rate for stable marriages and declines to a relatively low level. Thus, the complex duration pattern could just as well be explained by a combination of simpler patterns (i.e., effects of unobserved heterogeneity, Vaupel and Yashin, 1985).

Figure 4 illustrates some of the results from the analysis of FHS data on separation. The dashed line (labelled **) is based on a regression which includes a cubic polynomial of marriage duration to represent an increasing then decreasing duration pattern of separation risk (common to childless marriages, as well as marriages with children).

In contrast, the other three lines (labelled A, B and C: corresponding to childless, one child and two child couples, respectively) are based on the best fitting equation in which there are multiple duration variables: duration of the union (CLU years + married years), duration childless (0 after the 1st birth), age of the oldest child at home and age of the youngest child (may equal the oldest). To date, only one other study appears to have examined similar fertility interactions with the duration variable (Hoem and Hoem, 1988).

The complexity of Figure 4 arises from the fact that a line representing separation risks by marriage duration for a given marital/labour force/fertility history can change abruptly. One cause of an abrupt change might be the birth of an additional child. How large the change might be would depend on the number and spacing of previous births. Subsequently, there are further abrupt changes when there are no more children of pre-school age or when the last child leaves home (i.e., the Empty Nest). Figure 4 contains two lines that purport to describe childless marriages. It is quite clear, however, that the dashed line (**) represents the confounding of

FIGURE 4

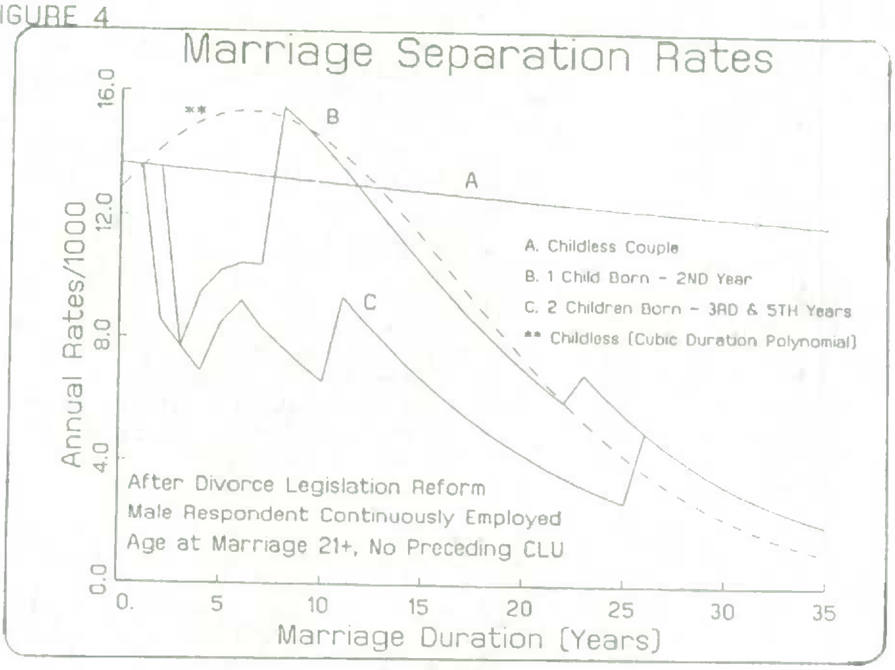


TABLE 4
Relative Risks of Separation by Work & Child Status

Child Status	Childless	1		
	Preschool Children (1+ aged ≤6)	0.78		
	School Age Children (all aged 7+)	1.19		
	Empty Nest	5.87		
Concurrent Work Status	Males	Females		
Working	1	1.86		
Not Working	1.63	0.68		
Never Worked	0.68	0.81		
Starting Work	0.94	2.09		
Stopping Work	2.75	1.28		
Non Response	0.79	0.83		
Work History (Since Marriage)				
	Childless		1+ Children Everborn	
	Males	Females	Males	Females
No Work Interruption	1	0.57	1	1
Work Interruption	1.10	0.68	0.61	0.76
No Work Since Marriage	1.74	0.54	1	1
Duration Coefficients (Log Hazard Slopes)				
Duration of Union (Married+CLU Years)	-0.1065			
Duration Childless (0 after 1st birth)	0.1022			
Age of Oldest Child at Home (Log)	0.4508			
Age of Youngest Child at Home (Log)	-0.1760			

duration effects on separation risk with the timing of marital fertility. The best fitting equation implies that separation risks generally decline with duration (even for childless couples A), although risks may be elevated as children reach critical ages. The presence of children is associated with both increases and decreases in separation risks.

Table 4 provides the relative risks associated with the interaction of work and child status in the best fitting regression. The most notable finding is that non-traditional work patterns (e.g., male not working, or female working, especially with children at home) are associated with elevated risks of marriage separation.

The results in Figure 4 and Table 4 are not consistent with either of the interpretations that it is normal for marriages to "wear out", or that many marriages are unstable at the outset. Rather, the risks of marriage separation normally decrease with duration, but those risks are also influenced by the numbers and ages of children. There is no need to invoke unobserved heterogeneity to explain episodes of elevated risk, where, for example, these may be associated with the youngest child attaining school age and/or the mother re-entering the labour force.

5. DISCUSSION

Two examples of analysis of FHS data have been presented here. In one case, changes in labour force status have been shown to influence marriage and CLU formation rates. By implication, factors influencing the labour market would contribute to an explanation of historical trends in marriage rates and average age at marriage. Similarly, a relationship has been demonstrated between risk of separation, on one hand, and marital fertility/labour force status, on the other. Therefore, historical trends in divorce will, in part, reflect trends in fertility and in male and female work patterns.

More important than the specific examples presented is the implication that the dynamics of marital, fertility and labour force decision making must be considered very carefully when modelling marriage rates, if we wish our models to reflect behaviour in a meaningfull way. Giving close attention to the dynamics also provides insight into which questions may or may not be directly addressed with available data. For example, it is clear from the results presented in section 3.2 that trying to determine the proportion of marriages that will eventually end in divorce is an extremely difficult problem. The source of the difficulty is that risk of divorce depends on dynamic work patterns involving work interruption and on the number and spacing of children.

The regressions on which this paper was based have been incorporated into a demographic microsimulation model (DEMOGEN), which integrates marital status and labour force models based on FHS data. Construction of a microsimulation model appears to be the only way of directly integrating and assessing the implications of complex event history models (e.g., with marital transitions contingent on labour force status, and labour force transitions contingent on marital status). Microsimulation is, moreover, the only direct way of estimating statistics like the proportion of marriages ending in divorce within the framework of the models.

NOTES

1. The permanence of separation (i.e., with no subsequent reconciliation) can not be established until legal divorce (or death), hence data on separation overstates marital breakdown. Similarly, for religious or other reasons some separated couples may never divorce, hence data on divorce understates marital breakdown. Nevertheless, the date of separation more accurately reflects the timing of marital breakdown than does the date of divorce.

2. The use of education level attained by the time of the survey instead of at the time of marriage results in greater difficulties than merely a confusion of time reference. Selection bias will result because person years at risk of marriage at lower levels of educational attainment exclude early years of education for individuals who eventually attain a high level. The bias would typically inflate marriage rates associated with low attainment. Similar selection biases might be associated with variables like religious activity (i.e., at the time of the survey) which are subject to change.

3. Transitions from the initial statuses SNGL, CLU or SEP tend to occur within a relatively short time interval. Consequently, in these cases, it was possible to limit attention to recent person years at risk while maintaining an adequate sample size. The date after which divorce legislation reform became law (July, 1968) provided a convenient demarcation for what was to be regarded as recent. However, since marriages tend to be long lived, all possible person years at risk of separation were used (i.e., for transitions MAR → SEP). This was particularly important in attempting to estimate relative risks of separation after the last child had left home.

REFERENCES

Balakrishnan, T.R., Rao, K. Vaninadha, Lapierre-Adamcyk, Evelyne, and Krotki, Karol J. (1987), "A Hazard Model Analysis of the Covariates of Marriage Dissolution in Canada," Demography, 24(3), 395-406.

Burch, Thomas K. (1985), Family History Survey: Preliminary Findings, Catalogue 99-955, Statistics Canada, Ottawa.

Burch, Thomas K. and Madan, Ashok K. (1986), Union Formation and Dissolution: Results from the 1984 Family History Survey, Catalogue 99-963, Statistics Canada, Ottawa.

Cox, D.R., and Oakes, D. (1984), Analysis of Survival Data, London: Chapman and Hall

Grenier, Gilles, Bloom, David E., and Howland, D. Juliet (1987), "An Analysis of First Marriage Patterns of Canadian Women," Canadian Studies in Population, 14(1), 47-68.

Hannan, Michael T., and Tuma, Nancy Brandon (1978), "Income and Independence Effects on Marital Dissolution: Results from the Seattle and Denver Income-Maintenance Experiments," American Journal of Sociology, 84(3), 611-633.

Hoem, Britta, and Hoem Jan M. (1988), "Dissolution in Sweden: The break-up of conjugal unions to Swedish women born 1936-60," Stockholm Research Reports in Demography 45, University of Stockholm, Section of Demography.

Hoem, Jan M. (1985), "Weighting, misclassification, and other issues in the analysis of survey samples of life histories," in Longitudinal Analysis of Labor Market Data, eds. James J. Heckman and Burton Singer, Cambridge: Cambridge University Press, 249-293.

____________ (1989), "Limitations of a Heterogeneity Technique: Selectivity Issues in Conjugal Union Disruption at Parity Zero in Contemporary Sweden," Stockholm Research Reports in Demography 56, University of Stockholm, Section of Demography.

Hogan, D. (1978), "The effects of demographic factors, family background and job achievement on age at marriage," Demography, 15(1), 155-175.

McCullagh, P., and Nelder, J.A. (1983), Generalized Linear Models, London: Chapman and Hall.

Teachman, Jay D., Polonko, Karen A., and Scanzoni, John (1987), "Demography of the Family," in Handbook of Marriage and the Family, eds. Marvin B. Sussman and Suzanne K. Steinmetz, New York: Plenum Press, 3-36.

Vaupel, James W., and Yashin, Anatoli I. (1985), Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics, American Statistician, 39(3), 176-185.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

TIME SERIES IN DEMOGRAPHY

R. Pressat[1]

Basic Elements of Demographic Time Series

Some basic time series used in demography include:

- series of vital events: births, marriages, divorces, deaths, ...
- series of population states.

These series can correspond to various intervals of time. While series of vital events are most frequently organized by year, they can, for lack of something better, be arranged by multi-year periods or still, for purposes of specific studies, by short periods (month, day, or even hour). As for population states, their periodic fluctuations are most often tied to the frequency of the census or to estimates made by statistical agencies during intercensal periods.

We wish to point out that our observations will not cover the case of events arranged for small intervals for the purpose of determining the frequency of the event considered for a given period (year, week, day).

Nature of Time Series

The previous series of basic data can only give rise to limited interpretation; they usually constitute the raw material from which time series of indices can be established. These lend themselves better to analysis. Among these indices, we cite those resulting from the computation of ratios or proportions:

- crude rates (mainly birth and death);
- specific rates of the age-specific type or, more generally, the duration-specific type (for example, marital fertility -- duration-specific rate);
- probabilities such as the probability of death, of marriage, ...
- rates in terms of ratios between two populations; the population shown in the numerator refers to a sub-population of the population in the denominator (for example, the proportion of single persons by age and sex at a given date, the labour force participation rate, the percentage of children attending school, ...).

Time Series as the Basis for Demographic Analysis

Demographic phenomena, like all social phenomena, are immersed in time and their analysis could not be done without reference to this dimension. In short, the production and analysis of time series are at the heart of demographic analysis.

These series can be analysed by themselves or in conjunction with other series to determine associations between the occurrence of demographic phenomena and that of other phenomena relating to the life of individuals in society.

First Dilemma: Transversal or Longitudinal Analysis?

When dealing with a series of age-specific rates, for example, or a series of summarized indices (cumulative average fertility rates), what type of analysis should be performed?

- In the area of mortality, it is customary to acknowledge the priority of transversal analysis for two reasons: the concern for keeping as close as possible to reality and thus detecting any negative trend in order to remedy it, and more especially, the conviction that the current situation is not affected by the past or is only slightly affected by it. However, this reasoning does not exclude the appropriateness of longitudinal syntheses in the presence of rapidly changing phenomena, as the transversal view may prove distorting. This type of distortion flows from a very general scheme synthesized in Figure 1. The full lines refer to generations at 10-year intervals, G_0, G_{10}, G_{20}, G_{30}, and represent the variation, with age, of the risk associated with a given phenomenon. In a and b, this risk, at the same age, increases with younger generations. However, in a, it increases with age in the same generation, and in b, it decreases; in c and d, the risk, at the same age, decreases with younger generations whereas in c, it increases with age in the same generation, and in d, it decreases. The transversal synthesis obtained by borrowing appropriate points from the generation lines (dotted lines) produces quite a different view of the phenomenon: it suggests a

[1] R. Pressat, Département de la Conjoncture, Institut national d'études démographiques, Paris, France.

decrease (in a) or an increase (in d) where increase and decrease with age persist respectively, whereas in b and c, the decrease and increase with age observed in the generations are amplified by a transversal view (it should be noted, however, that in addition to the previous schemes, other classifications can exist leading to less significant distortions).

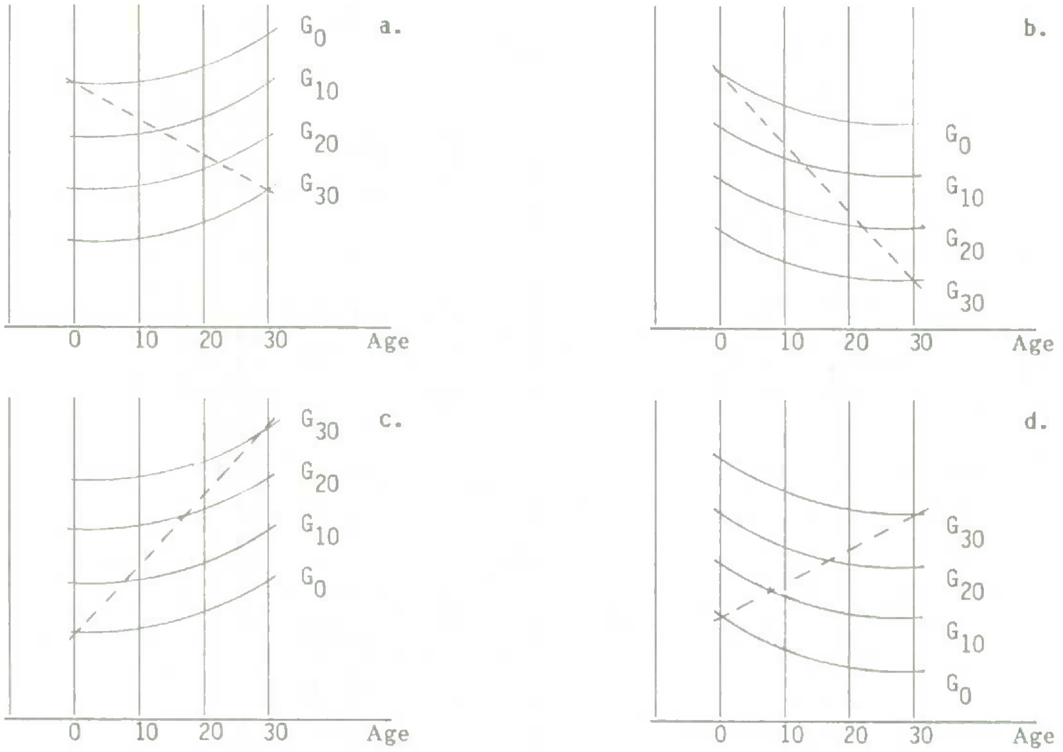


Figure 1

- In the area of fertility, the priority of a longitudinal view of the phenomenon is generally recognized because of the possible influence of the past on the current behaviour of generations. Without reconstructing the history of the fertile generations, it is possible, and sometimes advantageous, to look at the series of annual data which take into account past events in the reproductive life of women. Figure 2, which shows average age-specific fertility rates in France from 1950 to 1987, provides a perfect example of this. We will restrict our comments to the post-1964 period during which fertility decreased substantially. This period of sharp decrease has been identified as taking place between 1964 and 1976 (during this period, the total fertility rate fell from 2.90 to 1.83 births per woman). This decrease is observed without reprieve at all ages after age 28; before age 29 and towards the middle of the period, a levelling-off can be observed with an occasional variable increase. After 1976, the decrease is sustained before age 22; it hesitates between ages 22 and 25 and shows a tendency to increase beyond these ages with a sudden and generalized resurgence in 1983. One is naturally led to question the contrasting behaviour of the younger and older generations and this is when it is appropriate to consider their respective pasts: the recent increase in fertility for the approximate age group 27-37 is undoubtedly due to a change in schedule marked by substantially fewer births at the beginning of the reproductive period and, as a corollary, a slight increase towards the end of this period. This analysis also shows the promise of a stabilized total fertility rate in the next few years to the extent that the currently low fertility rate of the younger generations eventually levels itself off, and the partial recovery at the end of the reproductive history ensures compensation during the present transitional period preceding a future period of stability.

Time Series of Population States

These types of series are based on a series of censuses and they are of particular interest when the census frequency is regular, as the intercensal interval is equal to the range of the age categories. The French censuses were perfect examples of this for a long time as their five-year interval matched the decomposition of the population into five-year age groups; under these conditions, one has the benefit of a good follow-up of the evolution of the various groups of five generations over the years. Hajnal (and G. Mortara before him but less systematically) was the first to fully demonstrate the significance of age and sex-specific single person rates as a means of studying first marriages. These rates, which can be called non-married rates, have their formal equivalent in labour force participation rates, graduation rates for certain degrees, school attendance rates, etc. They represent, except for slight adjustment factors, the single persons in the first marriage table, the working population in the labour force participation table, etc. In the area of fertility, the information refers to the average range of the cumulative fertility (either the details of the various family

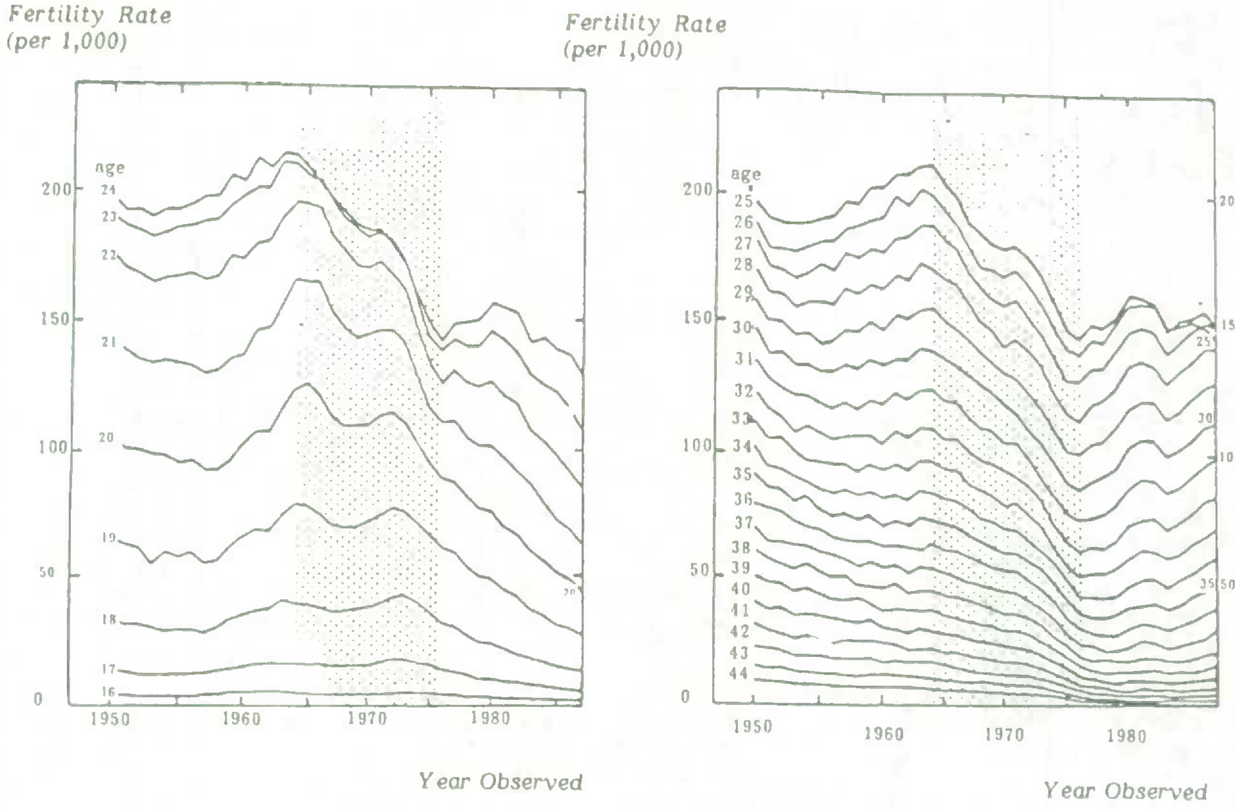


Figure 2 - France. Evolution of Average Age-Specific Fertility Rates

Source: G. Calot. Symposium on Population Change and European Society, European University Institute, Florence, December 7-10, 1988.

dimensions already achieved are known or only the average dimension is available); in this case, the information is included in the series of cumulative events in the fertility table.

Let us illustrate this by presenting the bachelor rate series derived from the long series of five-year French censuses dating from 1851 to 1946 (in 1956, 1961 and 1966, estimates supplement the missing censuses (Table 1). By limiting himself to certain typical generations, the reader may draw certain conclusions (Table 2).

Table 1 - France. Proportion of Bachelors (in %) in Various Generation Groups

	Generations												
Age Group	1821 1825	1826 1830	1831 1835	1836 1840	1841 1845	1846 1850	1851 1855	1856 1860	1861 1865	1866 1870	1871 1875	1876 1880	1881 1885
15-19 years	(99.7)	(99.7)	99.8	99.7	99.7	99.2	99.7	99.9	99.4	99.8	99.9	99.8	99.7
20-24 years	(89.0)	89.4	80.2	87.3	78.9	82.6	84.4	86.8	86.8	90.4	92.7	90.4	(87.0)
25-29 years	58.3	54.8	56.0	48.7	49.0	45.4	48.6	50.2	50.5	48.9	48.1	(46.0)	43.6
30-34 years	31.0	30.1	26.9	28.3	25.8	28.8	29.6	27.4	26.6	23.6	24.2	22.8	(26.0)
35-39 years	19.0	17.8	18.8	18.1	20.2	21.0	18.9	18.0	16.3	(16.8)	15.7	(16.5)	(15.6)
40-44 years	13.5	17.5	14.0	15.7	16.8	15.3	14.5	(12.7)	(12.9)	12.6	(12.8)	(12.4)	9.9
45-49 years	12.0	11.9	12.9	14.7	13.1	12.8	10.9	(11.3)	11.3	(11.3)	(10.6)	10.6	9.3

	Generations											
Age Group	1886 1890	1891 1895	1896 1900	1901 1905	1906 1910	1911 1915	1916 1920	1921 1925	1926 1930	1931 1935	1936 1940	1941 1945
15-19 years	(99.7)	(99.7)	(99.9)	99.4	99.4	99.4	99.5	(99.5)	99.2	99.4	99.4	99.3
20-24 years	(89.3)	(90.0)	(80.0)	76.4	78.0	76.3	(87.0)	80.6	76.9	77.4	78.5	-
25-29 years	(54.0)	(40.5)	34.7	36.0	35.7	(42.0)	46.0	36.2	32.6	34.5	-	-
30-34 years	(23.7)	20.9	18.7	20.2	(20.0)	26.1	21.0	18.1	17.6	-	-	-
35-39 years	10.6	12.7	13.2	(15.0)	16.8	15.6	12.9	13.1	(12.8)	-	-	-
40-44 years	10.2	10.3	(10.5)	12.5	13.0	11.0	10.5	(10.5)	(10.3)	-	-	-
45-49 years	8.7	(9.0)	9.6	10.8	10.7	10.1	(9.4)	(9.4)	(9.0)	-	-	-

Note: The figures in brackets are estimates.

Table 2: First-marriage Characteristics of Selected French Male Generations

Generations	Distribution of 1,000 First Marriages							First Marriage Before age 50	
	18-20 yrs	20-25 yrs	25-30 yrs	30-35 yrs	35-40 yrs	40-45 yrs	45-50 yrs	Number per 1,000 (18 yrs. old)	Average Age
1821-1825	22	253	365	225	79	39	17	890	28.7 yrs
1871-1875	28	317	389	172	55	28	11	900	28.0 yrs
1906-1910	39	439	355	89	39	28	11	900	26.3 yrs
1926-1930	38	470	339	88	38	16	11	915	25.9 yrs

Source: Jean-Claude Chasteland and Roland Pressat. La nuptualité des générations françaises depuis un siècle. Population, 1962, no. 2

Time Series and Mathematical Models

Attempts have been made to represent the development of demographic phenomena in time by mathematical functions. The first attempts were concerned with mortality (for example, the Gompertz and Gompertz-Makeham laws).

Formulation attempts concerned with other phenomena, such as fertility, took place much later, in 1931. At that time Wickell viewed them as a means of solving Lotka's equation. This requires the determination of the ultimate stable population obtained by maintaining stationary mortality and fertility conditions. At that time, as a means of verifying the adequacy of the model, reference was provided by the series of age-specific fertility rates for a given year (or period of several years).

How can the existence of time series, in this case the age-specific fertility rate series, contribute to the determination of a fertility function representative of the fertility of the population considered? A few comments on the nature of the mathematical functions most apt to represent age-specific fertility may answer this question.

In this regard, three types of density can be considered to represent these functions; all three belong to the K. Pearson family:

- the probability density associated with the gamma function, a particular Pearson type III case;
- the probability density associated with the beta function, a particular Pearson type I case;
- the polynomial form $(x-a)(\beta(x)^{2}$ also a particular Pearson type I case.

In these three expressions best suited to represent the fertility functions, the mathematical formula for fertility density may be expressed as follows:

$$f(x) : = D_{\beta}\gamma(x)$$

where D_{β} is the lifetime fertility (or total fertility rate in transversal analysis), and $\gamma(x)$ is the fertility schedule $(\int_{\alpha}^{\beta}\gamma(x)dx=1)$. Based on the structure of this formula, it appears that we are unable to arrive at a mathematical expression of the fertility density that expresses the relationship between the intensity and schedule of the phenomenon.

Time Series and Projections

"In the absence of laws that permit forecasting with certainty, the study of relationships between situations at various times is the only scientific basis for forecasting." "Forecasting the future consists of going from the present to the future based on knowledge of the past". (translated) These two quotes from a study by L. Henry and H. Gutierrez ("Qualité des prévisions démographiques à court terme". Population no.3, 1977) define aptly the nature of demographic forecasts and the place of time series analysis in the production of these forecasts.

In the following section we will distinguish the very short term forecast from the long term forecast.

Very Short Term Forecasts

The objective of a short term forecast is to determine the exact meaning of the most recent developments in a time series in the hope of detecting any reversal of trend.

In this regard, by following work carried out in France, we intend to test various forecasting methods retrospectively. By comparing results to subsequent evolution of time series, we will be able to make a judgment on the pertinence of these methods.

We will focus first on the problem of extrapolation of birth time series. In the short term, we will work on monthly series. These may be:

- the series of deseasonalized monthly values;
- 12-month moving sums.

Each of these series will be subjected to linear extrapolation by the method of least squares. The quality of these extrapolations will be evaluated on the basis of the values of the mean quadratic deviations between actual values and forecasts.

The forecasts thus produced may be distinguished by:

- the number of recent months taken into account, p;
- the length of the period considered, h;

the past period used being the period 1960-1974.

The results of this comparison between retrospective forecasts and actual values are shown in Figures 3, 4, 5 and 6.

We arrive at quite different conclusions depending on whether the monthly births of the last p months before the starting date of the forecast are used or whether the series of moving sums of births for the last p months is used.

Figures 3 and 4 highlight the importance of the past (p months) according to the forecast period (h). By considering data for only one month preceeding the starting point of the forecast (p=1) (forecasting then amounts to extending the last result observed), it can be seen that the inaccuracy of the forecast diminishes when the data for a larger historical period is used (however, the inaccuracy does increase for very long periods not shown on the graph). In short, the number of months used should not be too small in order to lessen the effect of random variations in the monthly series on the accuracy of the forecast. With the use of 12-month moving sums, we arrive at very different error profiles depending on the historical period used: an increase in this historical period leads to a rapid increase in error. In both cases, if a common historical period is used, as might be expected the error is all the more pronounced as the forecast period (h) increases. The error according to the forecast period is greater when moving sums are used (this comparison can be facilitated by dividing the ordinates of the graph in Figure 4 by 12).

Figures 5 and 6 give a different reading of the previous results:

- In Figure 5, it can be seen that the gain in accuracy is clearly a function of the length of the past period used for extrapolation; the simple extension of the value of the last month observed (p=1) is the most favourable procedure where the forecast covers 5 or more months;
- In Figure 6, the situation is slightly more complicated; while the length of the past period considered is an unfavourable factor, the simple extension of the most recent moving sum (p=1) is however the most judicious choice where the forecast covers 10 or more months.

Now, if we want to calculate the size of the errors resulting from this type of operation, we note the following:

- In the monthly series adjusted for seasonal variation, the margin of uncertainty, defined by the mean quadratic error, for approximately 70,000 monthly births, varies little with the forecast period and is in the order of 1,500 to 2,000 births, i.e. a relative error of 2% to 3.5%.
- In the moving sums series, the margin of uncertainty for approximately 840,000 annual births is 18,000 to 20,000 births depending on whether the forecast is for 1 month or 12 months, i.e. a relative error ranging from 2% to 2.4%.

The inertia of the deseasonalized series thus appears to be much greater than that of the moving sum series. confirming that it is better to base a one-month forecast on the former series (using p=1 according to Figure 5) than on the latter series (using p=2 as suggested in Figure 6).

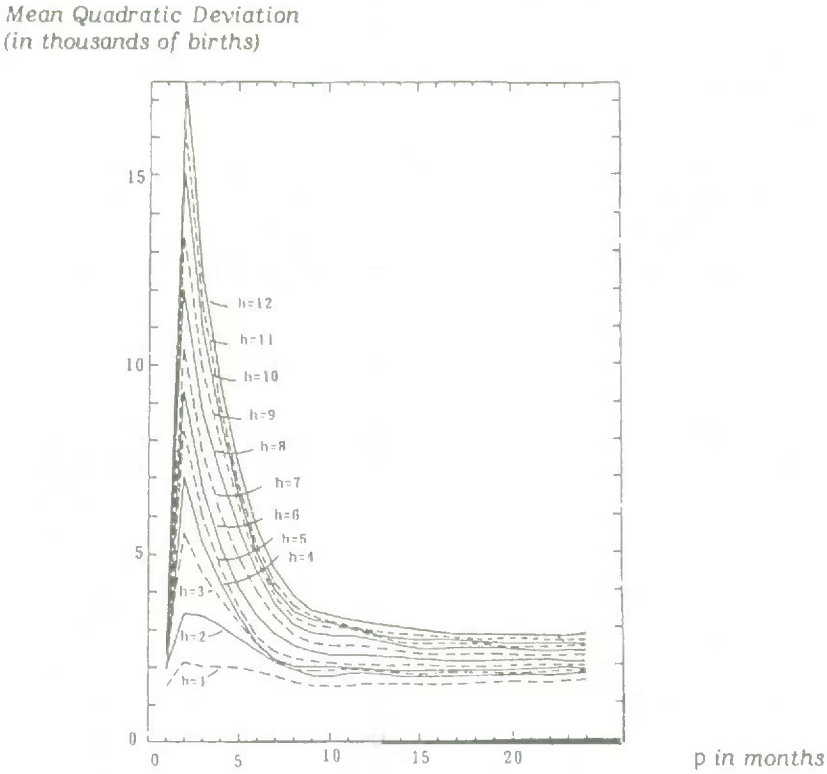


Figure 3: Mean quadratic deviation of the forecast of deseasonalized monthly births according to the number of months of historical data used, p and the forecast period, h.

Source: G. Calot and R. Nadot. Combien y aura-t-il de naissances dans l'année? Population, special issue, 1977.

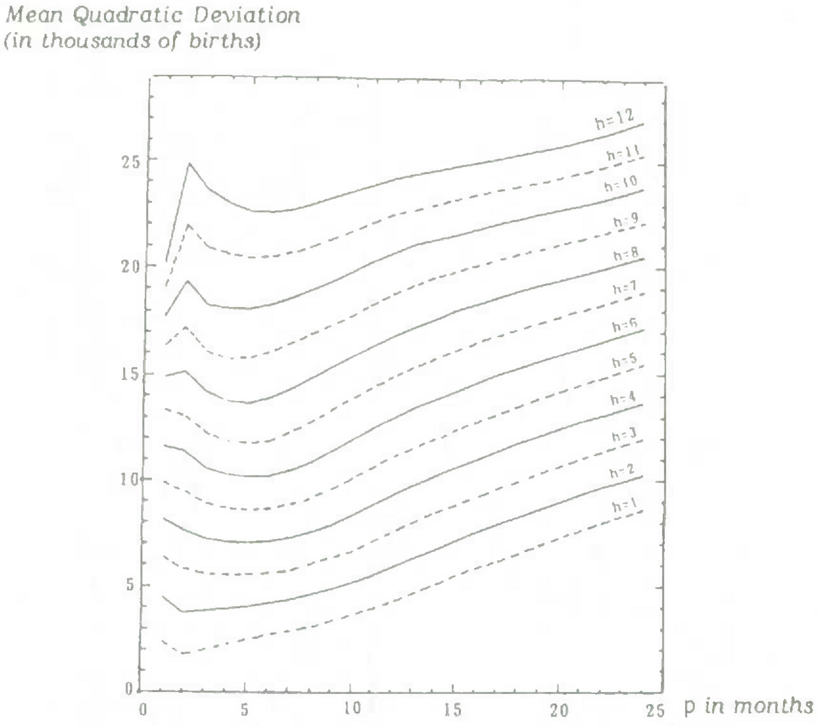


Figure 4: Mean Quadratic Deviation of the moving sums of births over 12 months according to the number of months of historical data used, p and the forecast period, h.

Source: G. Calot and R. Nadot. Combien y aura-t-il de naissances dans l'année? Population, special issue, 1977.

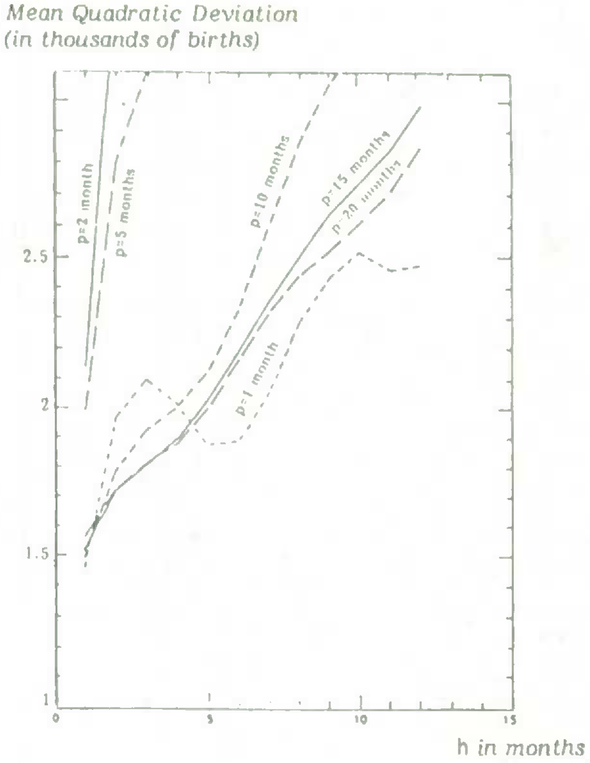


Figure 5: Figure 3 data at constant p

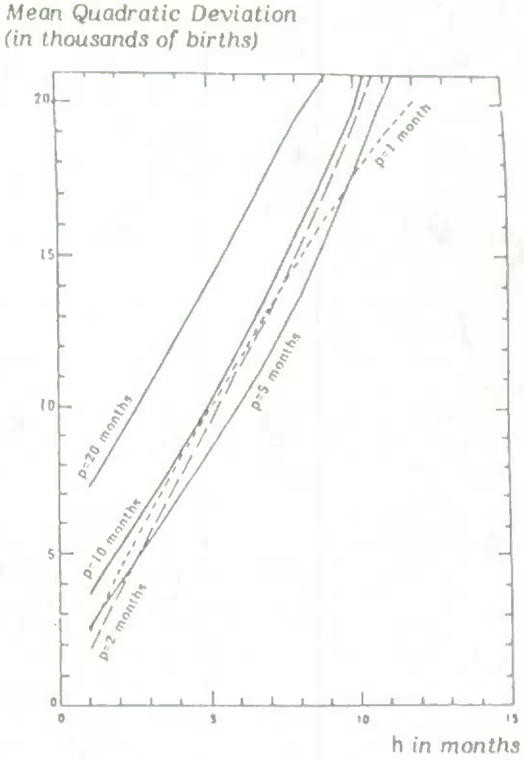


Figure 6: Figure 4 data at constant p

Forecasting the Total Fertility Rate

The use of the total fertility rate permits a better approach to analysing behaviours than the use of number of births alone. Better still, consideration of average age-specific fertility rates as being included in the computation of the total fertility rate leads to a more thorough analysis. We will specifically examine the problem of the use of time series in the projection of fertility rates.

In Figure 7, the rate at age x (attained), at date t, i.e. $f(x,t)$, corresponds to the shaded area. A first approach is the linear model:

$$f(x,t)=a(x)t+b(x). \tag{1}$$

This model does not take into consideration the cohorts' past (in this case the generations' past) as a factor likely to influence behaviour at a given age. To consider past fertility is to consider the cumulative fertility at the beginning of year t. This cumulative fertility may be expressed as follows:

$$\sum_{\xi=15}^{t-1} f(\xi,t-(x-\xi)).$$

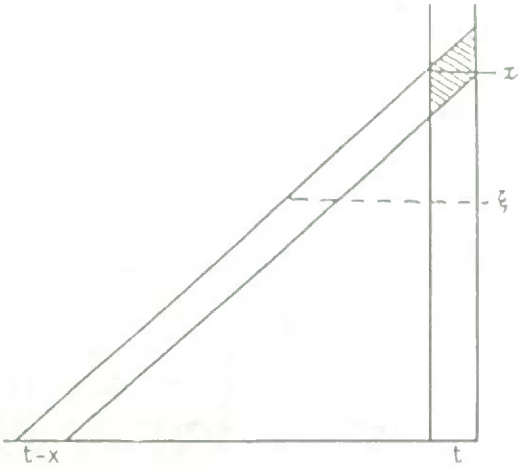


Figure 7

We can thus consider extrapolation using the following equation:

$$f(x,t) = a(x) \cdot \sum_{\xi=15}^{t-1} f[\xi, t-(x-\xi)] + b(x). \tag{2}$$

As indicated previously, the extrapolation can be based on a variable number of starting points; and, as previously, we have the feeling, a priori, that an optimal number of points exists that is sufficiently small to take recent trends into account but not so small that it would entail forecasts that would be strongly affected by random variations associated with an excessively small number of observations.

The study from which the previous results were extracted was conducted using French fertility and male and female first-marriage data for the period 1946-1982; the retrospective forecasts apply for the period 1977-1982, with extrapolations being made for single years (for example, 1977 is estimated on the basis of data not later than 1976, 1978 on the basis of data not later than 1977, etc.).

Figure 8 enables us to compare the two methods presented. The extrapolations done with the time series without taking into account previous generations are denoted by L and those done by taking into account past generations are denoted by C.

Interestingly, the best results are obtained for both nuptuality and fertility when past generations are excluded. Furthermore, with respect to first marriages, and as observed previously when we posed the problem relating to the projection of number of births, the best results are obtained when a limited number of previous points is considered (in this case 4). Finally, in the case of fertility it is of concern to note that the most satisfactory forecast is the simple extension of the value of the previous year.

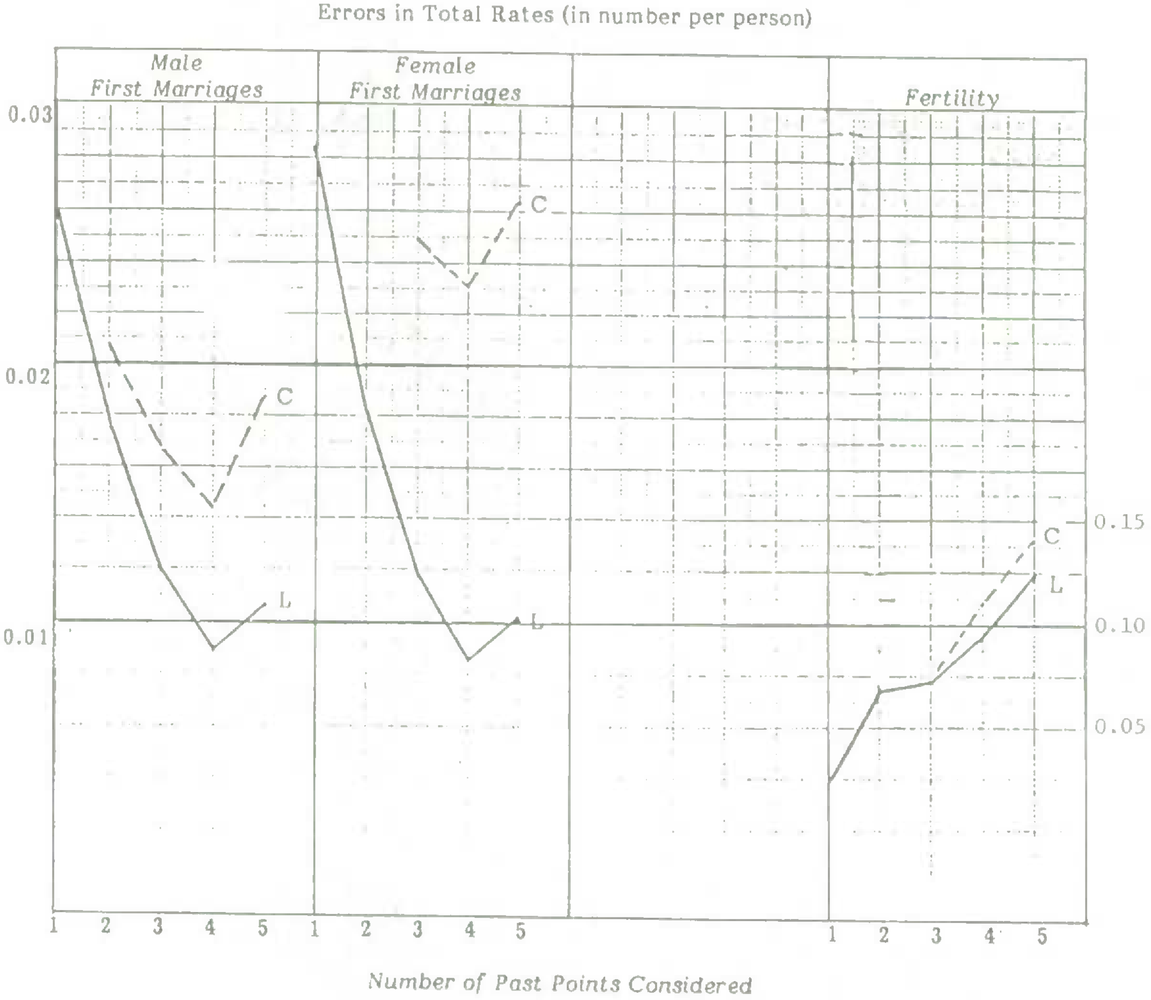


Figure 8: France. Single-year forecasts for the period 1977 to 1982 (average results).

Source: J.-P. Sardon. L'analyse démographique conjoncturelle: réflexions méthodologiques. Paper presented at XX^{ième} Congrès général de l'U.I.E.S.P., unpublished (Florence, 1985).

Considering that the best results are obtained using series of values for individual ages, observed over time, one might expect that better forecasts can be obtained using models that take into account both the previous generations and the trend at constant age. This amounts to completing the previous formula by the introduction of a linear term in t:

$$f(x,t) = a(x) \cdot \sum_{\xi=15}^{t-1} f[\xi,t-(x-\xi)] + b(x)t+c(x). \qquad (3)$$

However, we can modify the last two formulas by taking only recent generations into account, that is, by summing fertility starting not much before time t; in particular using p years of data. We are led to extrapolations based on the following formulas:

$$f(x,t) = a(x) \cdot \sum_{\xi=t-p}^{t-1} f[\xi,t-(x-\xi)] + b(x) \qquad (4)$$

or

$$f(x,t) = a(x) \cdot \sum_{\xi=t-p}^{t-1} f[\xi,t-(x-\xi)] + b(x)t+c(x).$$

These last two formulas open the way to a considerable number of possibilities; since we cannot examine them all, we can give preference to those where p=1, in other words, to those for which the acquired knowledge pertaining to the previous year, for the generations concerned, is taken into account alone or in conjunction with the results of the previous year for the same age. The previous generation is, therefore, also taken into account (J.P. Sardon, 1986).

Two Types of Medium Term Forecasts

In the same spirit as before, but with reference to a very different type of forecast, we will demonstrate how the choice of weak assumptions in forecasting can lead to options that may appear simplistic but have certain operational qualities. We are referring to population forecasts for the French departments.

Given these difficulties and since forecasting is bound to continue, a simple method is available to forecasters: extending the rate of growth observed in the previous period to the next period, that is, the subject of the forecast. This approximate method, which was applied from 1831 to 1875 -- except for certain gaps due to various anomalous situations -- led to the results shown in Figure 9. This figure does not measure the deviation between forecasts and observations; instead, it gives a graphical representation of the relationships between the growth rates of each five-year period and the corresponding rates of the five-year period immediately following (in this way one notes the correspondence between the growth rates of the population for a given province between the period 1831-1836 and the period 1836-1841). If forecasting the population of a province for a five-year period by extending the growth rate of the last five-year period to the forecast starting point was perfectly correct, the graphs in Figure 9 would then be reduced to a diagonal line (equal growth on the x-axis -- past rate -- and on the y-axis -- future rate).

However, this is not the case. The full line refers to the median values for each period considered; it is situated below the 45° line for the first three periods and most of the time above for the fourth, indicating slowing growth in the first case and accelerated growth in the second. The location of the lines for the first and third quartiles gives an idea of the inaccuracy resulting from extension of the past rate as a means of forecasting; the deviations, in absolute values, are most often in the range of 2-4%.

As unsatisfactory as this result may be, extending the growth rate of the previous five-year period represents an improvement over simply maintaining the population figure of the previous census; and the graphs will confirm that, most often, errors occurred at least 3 times out of 4 using this method.

These negative results favour the production of extensive time series containing as much substance as possible. Only then will the method be tested and its results compared with reality to a sufficient extent as to improve forecasting methods; analysis and a priori reasoning cannot achieve this.

We will now illustrate another type of mean-term forecast that will underline the importance of time series. We are referring to population forecasts by marital status. This is another case where the demographic analysis behind the constitution and transformation of the various statuses is so complex that it cannot lead to operationally practical forecasting. In this case, another type of time series is considered, the first example being the series of age and sex-specific married rates. An examination of Figure 10 will enable us to keep our explanation short. The graph shows the proportions of married women in 1947, 1952 and 1957; the full lines that connect the points representing these proportions are, because of their regularity, extrapolated (broken lines) for the years 1962 and 1967. Here, the overlapping of the lines relative to the observations served as a guide.

This procedure can be repeated with the rates of other marital statuses (for example, widowers and widows, divorced persons, and common-law unions); it remains to adjust the individual results so that their total at a given age coincides with the population at that age (a proportional decrease will see to that).

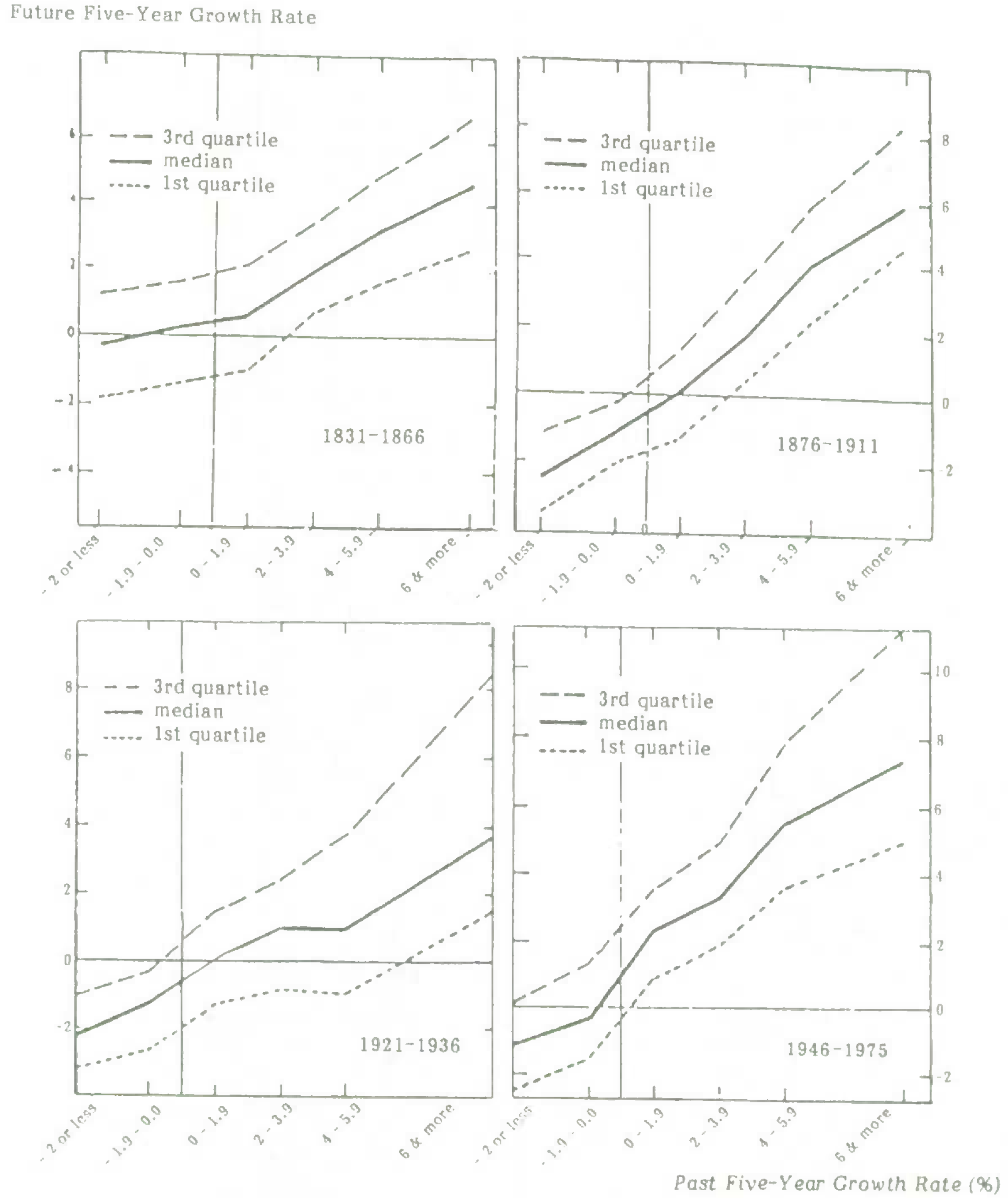


Figure 9: French Provinces. Medians and quartiles of the future 5-year growth rate (rate observed) following the past 5-year growth rate (rate forecast).

Source: L. Henry and H. Gutierrez. Qualité des prévisions démographiques à court terme. Population, no. 3, 1977.

Diversity of Time Series

In analytical problems as in forecasting problems, demographers often have to deal with several time series relating to the same phenomenon: are these various series equally valid as forecasting tools?

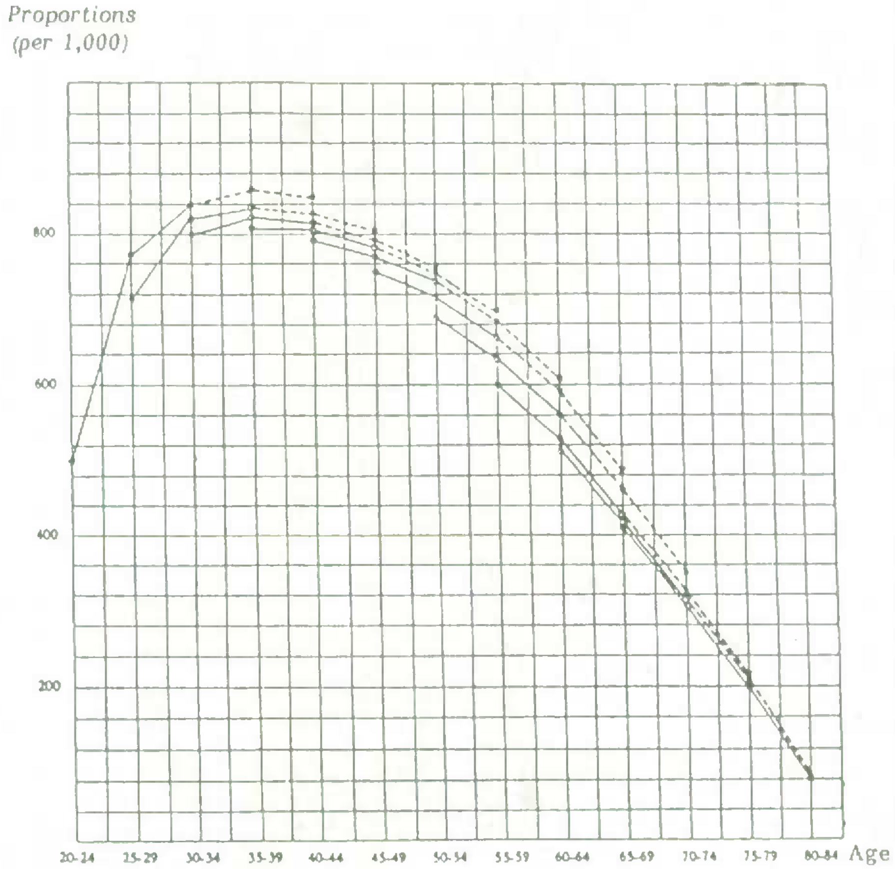


Figure 10: France. Proportions of women married in 1947, 1952 and 1957, and projections for 1962 and 1967.

Source: R. Pressat. Un essai de perspectives de ménages. Paper presented at Congrès général de l'U.I.E.S.P., Vienna (1959).

Diversity of Time Series

In analytical problems as in forecasting problems, demographers often have to deal with several time series relating to the same phenomenon: are these various series equally valid as forecasting tools?

The question does not arise in the analysis of past situations since the demographer is aware of the full meaning of each specific series. It is a different matter, however, when he must make a choice for purposes of forecasting. To illustrate this, we offer the following quasi-historic example which was the subject of one of our first studies under the supervision of L. Henry. This example has to do with the forecasting of legitimate births with the assumption of stable future indices, namely:

I. - average age-specific fertility rate (with a fixed illegitimate birth rate);

II. - fertility rate by age at marriage and marriage duration;

III. - fertility rate by marriage duration alone;

IV. - fertility rate by number of children born from current marriage and interval since eithe the marriage or the last birth. Two series IV_a and IV_b are derived from this approach taking into account the different birth schedules used.

V. - fertility rate of recent households (calculated in 1950), observed for duration of 0-11 years and extrapolated for future periods.

Figure 11 shows the extreme diversity of results. The indices were very unstable at the time undoubtedly because of the recent period of upheaval and its after-effects. But this is almost always the case: indices are influenced by past history, each in its own way depending on its structure. As a result, the projections conceal slightly dissimilar assumptions under the cover of analogous language.

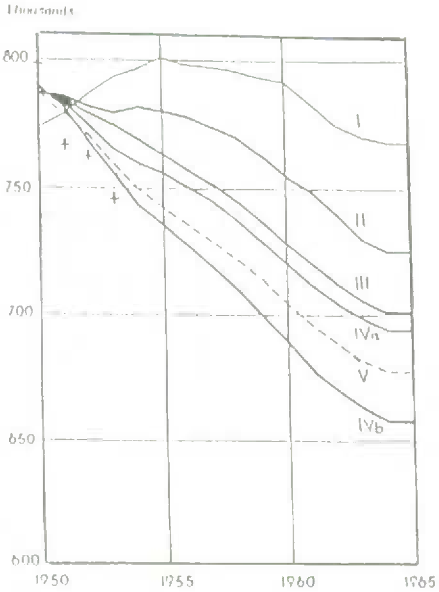


Figure 11: France. Forecast of legitimate births according to various methods.

Source: L. Henry. Perspectives de naissances après une perturbation de la natalité. Paper presented at Congrès de l'U.I.E.S.P., Rome (1955).

Concomitant Time Series

The comparison of time series relating to demographic events with other series relating to economic, political or social life is a key element of causal research in demography. This aspect of research does not utilize any specific demographic technique and we do not intend to examine this matter here. The findings of research in this area remain scant as is evident in the review carried out by H. Leridon in the area of fertility (see references).

General Observations

One may have the impression of a discussion touching many issues, with emphasis on the use of time series for forecasting. As mentioned in the introduction, time series are the raw material for population analysis and it must be recognized that there is a lack of information in this area due to poor data collection practices which have only recently become satisfactory. This lack of information has certainly had unfortunate consequences in the area of forecasting and that is why we stressed this aspect of the use of time series.

REFERENCES

Some of the previous discussion is supported by published results.

R. Pressat - *Distorsions introduites par une vision transversale des phénomènes en démographie.* Paper presented at Congrès de l'Institut international de statistique, Vienna, 1973.

G. Calot and R. Nadot *Combien y aura-t-il de naissances dans l'année?*, Population, special issue, 1977.

J.P. Sardon - *The short-term forecast of rates: Longitudinal or transversal perspective?* Materialen zur Bevölkerungswissenschaft, Heft 49, BIB, Wiesbaden, 1986.

L. Henrey et H. Gutierrez - *Qualité des prévisions démographiques à court terme.* Population, n° 3, 1977.

L. Henry - *Perspectives de naissances après une perturbation de la natalité.* Paper presented at Congrès général de l'U.I.E.S.P., Rome, 1954.

R. Pressat - *Un essai de perspectives de ménages.* Congrès international de la population, Vienna, 1959.

H. Leridon - *Natalité, saisons et conjoncture économique.* Cahier de Travaux et documents de l'INED, *n° 66, Paris, 1973.*

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

Inequality and Polarization: Is There a Disappearing Middle Class in Canada?

Michael C. Wolfson[1]

A. Introduction

There is a long standing and general interest in trends in income inequality in societies such as Canada. The popular wisdom is captured in the phrase "the rich are getting richer, and the poor are getting poorer". In the 1980s, a new twist in this interest emerged in the U.S. with the debate about "de-industrialization". Part of the hypothesis of de-industrialization is the gradual disappearance of relatively high paying blue collar industrial jobs. The view is that these jobs are being replaced by a mix of low paying, low skill service sector jobs ("McJobs") and a smaller number of high paying, highly skilled white collar jobs such as computer systems analysts.

This latter phenomenon was seen as being at the root of an emerging "disappearance of the middle class". A spate of articles has appeared on this, with Kuttner (1983) being an early example. More recently, this literature is reviewed in Loveman and Tilley (1988), with Canadian contributions by the Economic Council (1987), Myles (1987), Leckie (1988), and Picot et al. (1990). Many of the analyses in this area have been characterized by considerable heat as well light largely because of differing choices of concepts and definitions.

The objective of this paper is to set out the basic facts regarding trends in Canada over the past two decades for inequality and polarization. Polarization is our concept designed to capture the notion of the disappearing middle class, and is distinct from inequality as was pointed out by Love and Wolfson (1976).

Before considering the empirical results, we begin with a number of definitions and concepts. To anticipate the conclusions, we find no significant trends in income inequality, the same "stasis amid change" as found in Wolfson (1986a). However, we do find evidence of increasing polarization of the income distribution, but not for the reasons initially thought.

B. Inequality and Polarization of What for Whom

Most analysis of trends in income inequality tends to consider the total or after-tax incomes of families. In contrast, many of the analyses of the "disappearing middle class" examine the wage earnings of workers, and sometimes occupational titles, for example the relative growth rates of the highest and lowest paid occupations (e.g. Rosenthal, 1985). Thus, discussions of inequality and polarization can become confused if varying units of analysis and measures of economic position are used from one study to the next.

In this analysis, both total and disposable income, and income from wages and salaries only ("labour" income) will be considered. Also, both families generally and individuals with "non-trivial" labour force attachment will be considered. In addition, adjustments to family income to take account of the changing average size and composition of families will be examined. Thus, a variety of perspectives will be applied in a consistent way to analyzing trends in both inequality and polarization.

Another source of confusion in analyses of trends in inequality and polarization is the variety of statistical measures used in different studies. Sometimes the text in a study may be referring to inequality while the statistics in the accompanying tables are mathematically inconsistent with this concept. In other cases, the text refers to the disappearing middle class while the tables show measures of inequality. These confusions are of fundamental importance.

For our purposes, we shall use two broad groups of statistics from the income distributions to be analyzed -- those pertaining to inequality, and those pertaining to polarization. While it is still not widely recognized, these are distinct concepts. It is possible for one income distribution to be more equal than another, while at the same time showing greater polarization. Intuitively, inequality relates to the range of differences amongst the entire population, while the concept of polarization reflects the extent to which individuals or families tend to cluster in two distinct groups along the income spectrum.

We have chosen to use statistics to indicate trends in inequality and polarization that are mathematically straightforward in order to make the results as clear and intuitive as possible. However, the underlying reasoning for the specific choice of indicators is somewhat more technical, and is thus given in the Annex.

C. Overall Results

Table 1 presents the basic trends in inequality and polarization over almost two decades in Canada from a family perspective. Table 2 is identical in format, and is focused on individuals with "non-trivial" labour force attachment in each year. More precisely, Table 1 examines census families and their total income. Table 2, in contrast, considers only individuals age 15 or over who received labour income in the year in an amount greater than 2.5% of the average wage in the year. We refer to such individuals as "effective labour force participants" or ELFPs. Roughly speaking, they must have worked at least one week full-time at a rate of pay equal to the average wage, or at least two weeks full-time at the minimum wage.

These two tables thus represent two broad perspectives on the distribution of income -- the first focusing on families and their income from all sources, not just from working, and the second focusing on ELFPs and their income from work (including self-employment).

The data in all cases are drawn from special analyses of the Surveys of Consumer Finances. The specific years examined were chosen to provide the longest possible historical series for which the data are available and consistent, sample sizes are large, and the economy was at roughly similar points in the business cycle. These are exactly the same underlying data as have been used by Picot et al. (1990). This and the Picot et al. analysis are complementary, because the latter focuses on the subset of ELFPs who worked full-time and full-year. The ELFP data presented here underlie those used by the Economic Council of Canada (1989).

Each table has three groups of statistics. The first two rows show the average or mid-point of the income distribution in each year -- the mean and median incomes both expressed in constant 1986 dollars. While the figures have been rounded off to the nearest $50, sampling variability is such that the figures are really only accurate to about the nearest $500. From both the family/total income and individual/labour income perspectives, incomes grew most rapidly in the late 1960s and early 1970s. Growth in average incomes then slowed and over the early/mid 1980s stagnated and even declined.

The second group of figures pertain to income inequality. These are the shares of income accruing to each quintile group along the income spectrum, and the Gini coefficient, an overall index of inequality. From a family/total income perspective, all of these statistics show some variation from one year to the next. However, rough estimates of the sampling variability of these figures suggest that there are no statistically significant trends. For example, while the share of total income accruing to the top fifth of families varied by as much as 1.5%, the 95% confidence interval is probably at least two percentage points. Love and Wolfson (1976, appendix 2) in a similar context estimate the relative standard error of the Gini to be from 1.5 to 3.5% (i.e. values of 0.6 to 1.4%), depending on the size of the sample. Thus the differences in the Gini's of at most 1.5% (41.3 - 39.8) is unlikely to be statistically significant.

[1] M.C. Wolfson, Analytical Studies Branch, Statistics Canada, Ottawa, Ontario K1A 0T6

The question of a trend in individual labour income inequality is more borderline. The figures do point to an increase in inequality that could well be statistically significant.

In contrast with the absence of a trend in inequality, the data do show a clear trend toward increased polarization -- measured by a decline in the number of families and individuals with "near middle level" incomes. For example, Table 1 shows a decline of about one-seventh in the number of middle income families from 37.2% to 31.7% with incomes between three-quarters and one and one-half times the median family income.

Table 2 shows an even sharper decline of about one-fifth in the number of workers with middle level labour income, from 39.3% to 30.8% with incomes between three-quarters and one and one-half times median individual labour income in the bottom line. Probing a bit more into the figures, this decline occurred in roughly equal proportions both in the group with incomes in the range 75 to 125% of the median, and in the range 125 to 150% of the median. Of the 8.5 percentage point decline from 1967 to 1986 in the number of workers with incomes in the combined 75 to 150% of median income range shown Table 2, 5.4% "moved into" (not literally, since the data are not longitudinal) the greater than 150% of median income range, while the rest moved into the range with incomes of less than half the median.

These trends in polarization are almost certainly statistically significant both for individuals/labour income and families/total income. It is a practical demonstration that trends in polarization need not correspond to trends in income inequality; these are indeed distinct concepts.

While the statistics in Tables 1 and 2 have boiled down the concepts of middle income, inequality and polarization to 14 figures, this is still too many for purposes of more detailed analysis. Thus, for purposes of presentation, we shall focus on the Gini coefficient as our basic measure of inequality, and the share of the population with incomes between 75 and 150% of the median as our basic indicator of polarization. Still, the wider range of statistics has been examined in all the cases to be discussed to assure that the single summary statistic being displayed for each concept was accurately conveying the basic trends.

D. Possible Explanations -- Reporting Unit and Income Concept

One possibility is that the trends shown in Tables 1 and 2 above are some sort of statistical artifacts resulting from our particular choices of income reporting units and definitions of income. With one major exception, this section shows this not to be the case.

So far, we have focused on two income reporting units -- census families (CFs) which include parent(s) and never-married children living in the same dwelling as well as unattached individuals, and effective labour force participants (ELFPs). One other widely used broader definition of the family is economic families (EFs), defined as all related individuals living in the same dwelling. We can also define the subset of CFs that have at least one ELFP as a member, which we shall denote as ELFP-CF. Finally, we shall use Ind to denote the population of all individuals age 15+, whether or not they are ELFPs.

We have also focused so far on only two income concepts -- total or before-tax income, and labour income. We shall denote these BT and W income respectively. One other income definition of general interest is after-tax (AT) income. We would have made greater use of this concept except that, unfortunately, the 1967 data do not contain estimates of income tax paid.

Finally, for families we have generally used income per family. However, this is almost certainly a poor way of comparing the economic positions of families of different sizes. One way to account for differing family sizes is to use income divided by an equivalence scale. This is a scale of numerical factors that can be roughly interpreted as the relative income needs of families of different sizes.

The choice of equivalence scales is a matter of considerable controversy, for example as discussed in Wolfson and Evans (1989). We have chosen to use a scale that gives a weight of 1.0 to a single adult, 0.4 to second and subsequent adults in the family, and 0.3 to children (except first children in lone parent families who are given a weight of 0.4). These weights represent an equivalent adult unit scale or EAUs. Thus, a married couple with two children has a weight of 2.0.

Given such EAU scales, we can analyze distributions of family income where income is first divided by the number of EAUs in the family. Thus a married couple with two children and a total income of $25,000 would be treated as one family with an income per EAU of $12,500.

This kind of EAU adjustment might be expected to have an important effect because of the significant trends over the past two decades of declining fertility and increasing divorce rates, and thus a declining average family size.

Graphs 1.a to 1.d show the sensitivity of the basic results in Tables 1 and 2 to the various choices of income reporting units and income concepts just defined. Graph 1.a shows the trends in average real family income for four alternatives. Highest average total incomes are for EFs, not surprisingly because they have more members and thus more income recipients. Among CFs, after-tax income averages about $5,000 lower than before-tax income. Finally, average total income per EAU for CFs was about 60% of average total income per CF. However, irrespective of the income concept or the reporting unit definition, the general historical pattern is the same as noted in connection with Tables 1 and 2 -- substantial real income growth in the late 1960s and early 1970s, but stagnation and decline in the early/mid 1980s.

Graph 1.b shows trends in real average incomes for individuals. The figures in Table 2, it may be recalled, were for ELFPs and their labour (W) income. The ELFP,BT curve shows average total before tax income for these same individuals; it is higher by a few thousand dollars. The top line shows average total labour income for CFs with at least one ELFP. These amounts are 50 to 75% higher than average ELFP,W incomes because the earnings of spouses and children have been aggregated. While the ELFP,W and ELFP,BT curves show an almost identical time trend, the ELFP-CF,W curve shows steeper growth up to 1981. This is almost certainly the result of increasing female labour force participation rates over the period. However, this trend was not sufficient to prevent stagnation in average labour incomes among CFs with at least one ELFP over the 1981 to 1986 period.

Finally, the lowest curve in Graph 1.b (IND,BT) shows average total income among all individuals age 15+, not just ELFPs. It lies below the ELFP,BT curve because individuals who are not significantly attached to the labour force tend to have much lower incomes on average -- either none at all or modest amounts of investment income or income principally from government transfers. Still, as in Graph 1.a, the general time trends are consistent.

Graph 1.c focuses on inequality as measured by the Gini coefficient. The curves are so tightly clustered that only the top and bottom curves are labelled. The three non-labelled curves are all for total before tax income -- CF,BT; CF,BT/EAU; and ELFP,BT. While the choice of reporting unit and income concept has some effect on the measured level of inequality, it has negligible impact on the apparent trend. For the BT curves, there is no significant trend in inequality; it is generally constant as earlier concluded in Wolfson (1986a). For the W curves, there is a small upward trend, as discussed in conjunction with Table 2 above.

Finally Graph 1.d shows a variety of curves for trends in polarization, measured as the percentage of the population with incomes between 75 and 150% of the median. Again as in Tables 1 and 2, there is a clear downward trend in the proportion of middle income units, with some differences in levels corresponding to the different concepts.

There is one major exception, however. At the family level, the trend disappears when family total income is adjusted for variations in family size (the curve labelled CF,BT/EAU). In contrast, it is clear that income from working has become more polarized, whether we consider ELFPs individually (the curve labelled ELFP,W), or aggregated in census families (ELFP-CF,W). It also appears that total income of families (CF,BT) has become more polarized, as also shown in Table 1 above.

The implication is thus that declines in family size have been associated with changes in family incomes in a way that is offsetting from the viewpoint of polarization. The "disappearance of the middle class" from a family perspective using total income is apparently an artifact due to the failure to take account of systematic changes in family size.

On the other hand, the trends with respect to polarization at the individual level remain clear, and are associated with changes in the labour market.

E. Possible Explanations -- Labour Income Polarization

We turn now to an examination of other factors that might account for the increased polarization of workers' labour income. The "story" that is often told about the disappearing middle class relates to concepts like "de-industrialization" and "de-skilling". Unfortunately though, the available data in the Surveys of Consumer Finances being drawn upon for this analysis are not well suited to assessing such concepts. Thus, we shall be confined to more conventional variables -- age, sex, full- or part-time, industry, and occupation. Moreover, for the latter two variables, we shall have to make do with some quite coarse classifications due to limitations in the data.

For each group of variables the approach will be the same. The population of ELFP individuals will be divided into 4 to 6 mutually exclusive groups, and then four graphs will be examined. All have calendar year along the horizontal axis, exactly as in Graphs 1.a to 1.d above. The first graph shows the proportional distribution of the population among the various groups and how it has evolved over time. The second shows how the average income of each group compares to the overall mean income, relative mean income, expressed in percent. The last two graphs show the same inequality and polarization measures as before -- the Gini coefficient and the share of the given population with incomes between 75 and 150% of their median.

Age Structure Graphs 2.a to 2.d show the trends in the four variables just described for each of four age groups: 15-24, 25-34, 35-49, and 50+. The oldest age group declines over the two decades as a proportion of the effective labour force, reflecting the declining participation of older males. Youth, on the other hand, rises a bit from 1967 to 1973, but declines thereafter. This probably reflects the sharp drop in fertility (the "baby bust") after 1966 showing up as slowing growth of this age group in the early 1980s, and the poor job prospects for youth in the 1980s discouraging labour force entry. The 25-34 group is growing as a proportion of the total work force reflecting the entry of the baby boom birth cohort, while the growth of the 35-49 age group is most likely due to increases in female labour force participation.

In terms of relative incomes in Graph 2.b, youth have the lowest levels at about half the average, while the 35-49 age group are the highest at about 125%. There are no very strong trends in relative mean incomes among the four age groups.

All four age groups show virtually parallel trends in income inequality, and the levels are fairly similar in Graph 2.c. Thus, changes in the age structure of the effective labour force are unlikely to account for the overall small upward trend in inequality.

Similarly, all four age groups show generally parallel trends in polarization in Graph 2.d, though the levels are quite different. Youth have the smallest proportion with middle level incomes, and the highest measured labour income inequality. The 25-34 age group is the opposite. The declining proportion of middle level incomes pervades all four age groups, so again the overall trend in polarization observed earlier cannot be accounted for by changes in age structure.

Sex and Full- or Part-Time Status We turn next to a different four-way disaggregation of the ELFP population in Graphs 3.a to 3.d. This time, the population has been divided by sex, and whether or not the individual worked full-week (more than 35 hours per week usually) and full year (50+ weeks). If either of these conditions was not met, the individual was classified as part-time. The rise in female participation is clearly evident in Graph 3.a. Moreover, for both males and females, part-time workers represent an increasing proportion of the working population.

The declining proportion of male full-time workers, and increasing proportion of part-time workers was associated with an increase in the relative mean income of full-time males, and relatively speaking, and even larger increase in the relative mean earnings of full-time females, as shown in Graph 3.b. While female part-time workers increased both in proportion and in their relative mean earnings, albeit from a low starting point, their male counterparts experienced a decline in relative earnings in the early/mid 1980s.

Unlike the disaggregation by age in the previous set of graphs, for this set of groups the trends in within-group inequality shown in Graph 3.c are small relative to the differences between groups, particularly full- and part-time. Thus, shifts in the composition of the population can account for some of the overall trend in labour income inequality. This trend is small but it is upward, while within-group inequality is trending downward in the first time period, and is mixed in the second. Thus, the upward overall trend in inequality must be at least partly attributable to a shift in the work force to part-timers who have both lower and more unequally distributed earnings.

Similarly, these compositional changes in the labour force appear to account for some of the trend in polarization. Within the male and female part-time groups, there is virtually no trend in polarization; nor is there any very pronounced trend for full-time males as shown in Graph 3.d. The only clear trend is among full-time females, and then only over the latter two time periods. Thus, the increasing proportion of the effective labour force that is engaged part-time would appear to account for some portion of the overall trend toward increased polarization of labour incomes.

Industry Graphs 4.a to 4.d give corresponding results by broad industry groupings. These groupings are the best that can be defined in a consistent manner across the four surveys, and caution should be exercised because of the large number of "uncoded" industries in the 1967 and 1973 data. Not surprisingly, the largest growth has been in the services sector (wholesale and retail trade, personal and business services, finance, distribution), while the main area of decline (aside from "uncoded") was in manufacturing. However, leaving aside the smallest groups, primary (agriculture, forestry, fishing, mining) and construction, there were no significant trends in relative mean earnings, nor were there major differences in levels.

All industry groups show upward trends in labour income inequality, and all show stronger downward trends in the proportions with middle level incomes. Thus, even though there have been substantial shifts in the work force between sectors from 1967 to 1986, particularly from manufacturing into services, these shifts do not help to explain the increase in polarization. The reason, simply, is that polarization increased substantially within each of the major industrial sectors.

Occupation Finally, Graphs 5.a to 5.d show the corresponding trends for occupational groups. As in the industrial classification, this one is quite coarse, and includes a large number of "uncoded" occupations in 1967 and 1973. Again as might be expected, the major decline has been in the blue collar occupations, while the largest increase has been in low skilled/low paying white collar jobs ("lo-white" -- clerical, sales, service; "hi-white" is professional and technical). Aside from "blips" in the trends for the relatively small management group, which might be no more than an artifact of changes in occupational coding from 1967 to 1973, the trends in inequality and polarization within occupational groups are generally parallel. Thus, even though the occupational composition of the labour force has changed considerably, it cannot be used to explain the "disappearing middle" of the Canadian work force, at least for the broad categories being used.

F. Conclusions

A considerable literature has accumulated in the U.S. concerned with the "disappearing middle class". This literature has tended to explain the phenomenon in terms of "de-industrialization" and "de-skilling", the erosion of well-paying blue collar jobs in manufacturing and their replacement by a mix of low skill service sector "McJobs" and, to a lesser extent, high skill high tech sector jobs. (The more recent Canadian literature is less polemical and more cautious in interpreting the data.)

The phenomenon of the declining middle has itself been observed or not observed in part due to the wide variety of statistical indicators that have been used. In turn, the literature has tended to confuse the concepts of inequality and polarization -- the latter being our term for the disappearing middle class.

This paper has assessed the existence of the phenomena in Canada from several perspectives, and sought to determine associated trends which may have played a determining role. Unlike income inequality among families, which has remained stable over the past two decades, there is evidence of increased polarization. Polarization has increased both from the perspective of families and their total income before-tax, and for individual workers and their labour income.

In the case of families, the increase in polarization appears associated with changes in family size and composition. These latter changes are mainly due to the decline in fertility, and the increase in divorce.

For individual workers, the main factors seem to be the increase in female labour force participation, and the increase in part-time (i.e. less than full-week or full-year) work. Other factors such as changes in the age, occupational, and industrial composition of the work force do not appear to account for the increase in polarization.

Of course, these results are tentative, particularly due to the limited detail, and in some cases quality of the older survey data. More definitive results must await analysis of census data.

Reference

Atkinson, A.B. (1970), "On the Measurement of Inequality", *Journal of Economic Theory*, Vol. 2.

Beach, C.M. (1988), "The 'Vanishing' Middle Class?: Evidence and Explanations", *Queen's Papers in Industrial Relations*, Industrial Relations Centre, Queen's University at Kingston.

Cowell, F.A. (1977), *Measuring Inequality*, Oxford, Philip Allan Publishers.

Economic Council of Canada (1987), *Innovation and Jobs in Canada*, Catalogue No. EC22-141/1987E, Canadian Government Publishing Centre, Supply and Services Canada, Ottawa.

Economic Council of Canada (1990), *Good Jobs, Bad Jobs - Employment in the Service Economy*, Catalogue No. EC22-164/1990E, Canadian Government Publishing Centre, Supply and Services Canada, Ottawa.

Harrison, B., C. Tilly, and B. Bluestone (1986), "Wage Inequality Takes a Great U-Turn", *Challenge*, March-April.

Kuttner, B. (1983), "The Declining Middle", *The Atlantic Monthly*, July.

Leckie, N. (1988), *The Declining Middle and Technological Change: Trends in the Distribution of Employment Income in Canada, 1971-84*, Discussion Paper No. 342, Economic Council of Canada, Ottawa, January.

Levy, F. (1987), "Changes in the Distribution of American Family Incomes, 1947 to 1984", *Science*, Vol. 236, pp. 923-27, May.

Love, R. and M.C. Wolfson (1976), *Income Inequality: Statistical Methodology and Canadian Illustrations*, Catalogue 13-559 Occasional, Statistics Canada, Ottawa, March.

Loveman , G.W. and C. Tilly (1988), "Good Jobs or Bad Jobs: What Does the Evidence Say", *New England Economic Review*, January/February.

Myles, J. (1988) "The Expanding Middle: Some'Canadian Evidence on the Deskilling Debate", *The Canadian Review of Sociology and Anthropology*, Vol. 25:3, pp. 335-364, August.

Picot, G., J. Myles and Ted Wannell (1990), *Good Jobs/Bad Jobs and the Declining Middle 1967-1986*, Analytical Studies Branch Research Paper Series No. 28, Statistics Canada, Ottawa.

Rosenthal, N.H. (1985), "The shrinking middle class: myth or reality?", *Monthly Labour Review*, March.

Wolfson, M.C. (1986a), "Stasis Amid Change -- Income Inequality in Canada 1965-1983", *Review of Income and Wealth*, December.

Wolfson (1986b), "Polarization, Inequality, and the Disappearing Middle*", Statistics Canada, Ottawa, Mimeo.

Wolfson, M.C. and J. Evans (1989), *Statistics Canada's Low Income Cut-Offs -- Methodological Concerns and Possibilities: A Discussion Paper*, Statistics Canada, Ottawa, December.

Annex

The Measurement of Inequality and Polarization

The most broadly accepted formalizations of the concept of economic inequality are all related to the Lorenz curve. This curve is a way of displaying any income distribution -- a set of data showing how many income recipients (whether family units or individuals) there were at various levels of income. The Lorenz curve is a graph showing the cumulative fraction of the population along the horizontal axis, and their cumulative share of income (or other measure of economic position) along the vertical axis, assuming the population had been ranked in increasing order of their incomes.

The ordering of two income distributions (e.g. for two points in time) according to whether the Lorenz curve of one income distribution lies at least somewhere above and nowhere below that of another is the "gold standard" in virtually all axiomatized foundations of inequality measures (Atkinson, 1970; Love and Wolfson, 1976; Cowell, 1977). The most commonly used measure of income (or other indicator of economic position) inequality is the Gini coefficient. It is fully consistent with the ranking of income distributions given by Lorenz curves.

However, the Gini coefficient is not the only summary index of inequality that is fully consistent with Lorenz rankings. Others include the Theil, Theil-Bernouilli, Atkinson, and Exponential measures, and the coefficient of variation. When the Lorenz ordering is ambiguous -- i.e. the Lorenz curves for two income distributions cross so neither clearly dominates the other, these measures will generally yield different rankings.

Other popular inequality statistics are based on quantiles, such as the shares of income accruing to population quintiles. Quintile income shares are better considered inequality indicators rather than inequality measures. While they are never inconsistent with Lorenz rankings, quintile shares can remain unchanged even though Lorenz rankings do change.

There are, in addition, statistics that are used in discussions of inequality even though they are not even consistent with Lorenz curve rankings. These include the variance of logarithms of income (e.g. used in Harrison et al., 1986) and the inter-quartile ratio. Such statistics should never be used in these contexts because they simply do not measure what they purport to measure.

However, a more fundamental problem in the context of discussions of the disappearing middle class is that inequality measures, even if they are perfectly consistent with the Lorenz criterion, may be inconsistent with the desired concept. This point is illustrated in Figure 1, which shows two simple hypothetical income distribution densities. The first is a uniform density over the range from income 0.25 to 1.75, shown by a dashed line..

The second density, shown by a solid line, is clearly bi-modal, and has a somewhat depleted middle. We would argue that according to any sensible definition of polarization or disappearing middle, this latter density is the more polarized.

Is it also more unequal?

The answer is unequivocally no. The second density has been constructed such that according to **any** inequality measure that is consistent with the Lorenz criterion, it is more equal. This can be shown simply by the fact that the bi-modal distribution can be derived from the uniform distribution by two sets of progressive mean-preserving redistributive transfers in the sense of Atkinson (1970).

The first set of equalizing income transfers is from individuals in part p in the 0.75 to 1.00 part of the income range to individuals in part q in the lowest part, 0.25 to 0.50. The p's give the q's portions of their incomes equal to half the difference between their incomes -- 0.25 on average, so they move to parts p* and q* in the bi-modal distribution, in the 0.50 to 0.75 income range. Similarly, individuals in the highest part of the income distribution with incomes between 1.50 and 1.75, part m, give an average of 0.25 of their income to group n individuals in the upper-middle part of the distribution, incomes from 1.00 to 1.25. As a result of this set of progressive transfers, they both end up in the same 1.25 to 1.50 income range in parts m* and n* of the bi-modal distribution.

Thus, by construction, the bi-modal distribution is at the same time more polarized and more equal than the uniform distribution from which it was derived. Polarization and inequality are thus demonstrably different concepts, as first pointed out in Love and Wolfson (1976).

This result leaves open the question of what statistics should be used to measure polarization. In the literature on the disappearing middle, in addition to inequality measures, some authors have used quintile income shares, while others have used the fraction of the population in various income ranges defined in terms of the mean or median income. In fact, Figure 1 has been constructed in a particularly nasty way for these kinds of statistics.

Since the distribution is symmetric, the mean is equal to the median which is one. It can be shown that the share of the middle third of the bi-modal distribution is lower than the share of the middle third of the uniform distribution, while the share of the middle two-thirds rises in the transition to the bi-modal distribution. Thus, the income shares of various middle quantile groups are **not** necessarily consistent with any sensible formalization of the concept of polarization. In turn, this means that the large number of papers purporting to analyze the disappearance of the middle class which use inequality indicators such as quintile shares (e.g. Levy, 1987; Beach, 1988) are simply unable to detect the phenomenon they claim to be studying.

Moreover, the share of the population with "middle level incomes" goes up or down depending on how "middle" is defined in this example. This is easily seen by inspecting Figure 1. The population with incomes within 25% of the mean or median clearly falls, but the population with incomes within 50% of the mean or median rises. Thus, statistics that count the share of the population with "near middle" incomes are also **not** necessarily consistent with a sensible definition of polarization.

All is not lost, however. Given our improved formal understanding of the concept of polarization based on analysis of Figure 1, we can choose a set of statistics for detailed analysis. The population shares in various income ranges defined in terms of the median income, as in Tables 1 and 2 in the main text, are a good example. Then a summary statistical indicator like the share of the population with incomes in the range of 75 to 150% of the median can be used for purposes of discussion and graphical display, provided it is always checked by the analyst for consistency with the more detailed figures. This is what has been done in the analysis reported in the main text.

Furthermore, there are more complex formal approaches. These define a class of summary polarization indicators which are analogous to inequality measures, except that instead of being consistent with the Lorenz criterion, they are constructed to respond in a consistent manner to movements of a distribution toward bimodality. One such measure, WPOL, based on the Gini coefficient, is developed in Wolfson (1986b) and was used in the Economic Council (1987).

The WPOL measure has not been used here because it was considered more complex and less understandable than necessary for the analysis. The polarization measure that has been principally used, the population share with incomes between 75 and 150% of the median, while not necessarily formally consistent with the concept of polarization, is readily understandable and has been checked for consistency wherever it was used. The WPOL measure is formally consistent with the concept of polarization, and it could have been used instead. However, the general results would have been the same, and the reader would have had to learn about a new complex statistic to have felt comfortable with the results.

Still, if the topic of the disappearing middle class continues to gain in interest, it may prove beneficial to develop a more formal set of statistical measures such as WPOL for analytical purposes.

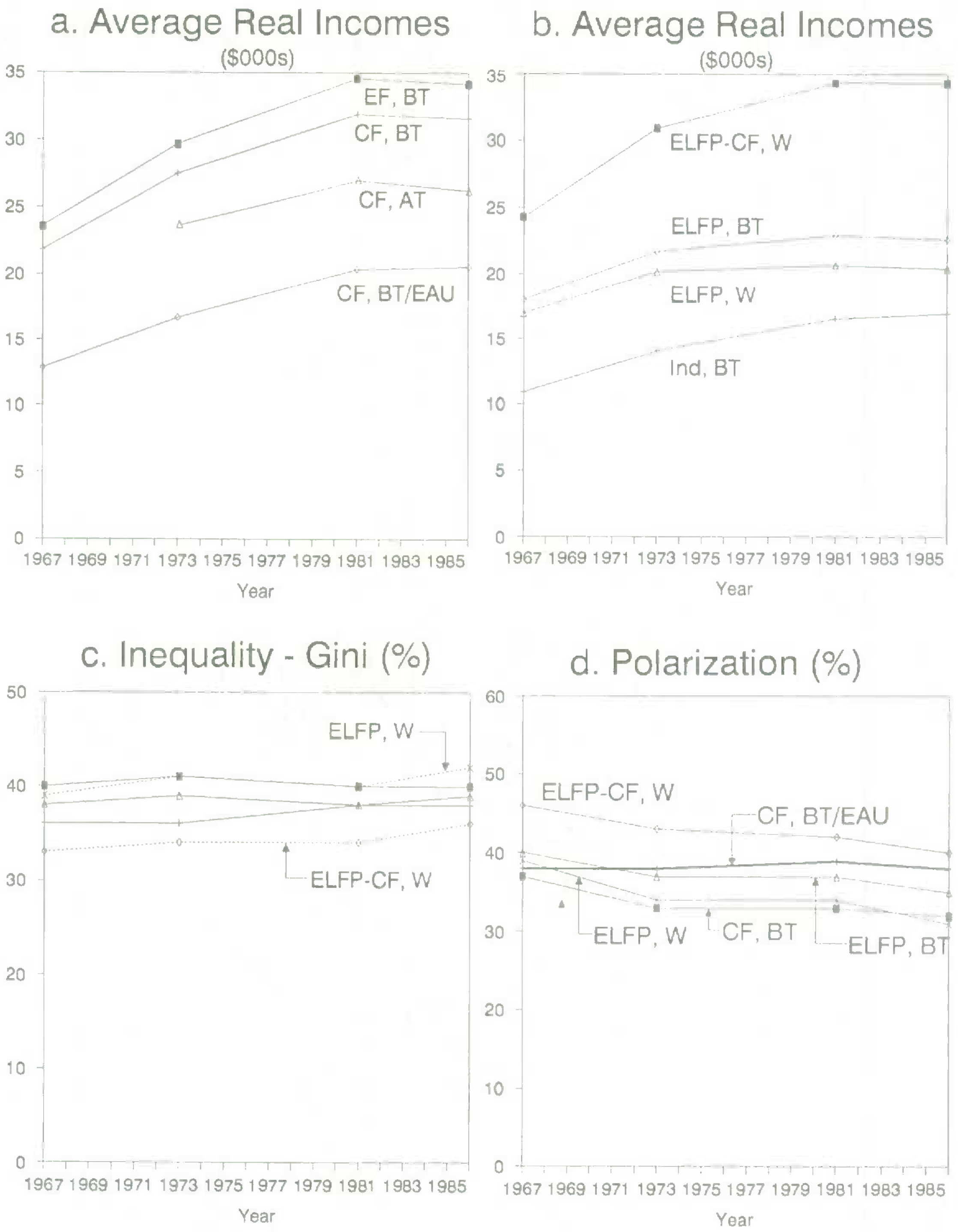
TABLE 1: Inequality and Polarization Statistics, All Census Families, Total Income

	Year			
	1967	1973	1981	1986
Mean income (1986 $)	21,850	27,550	31,900	31,650
Median Income (1986 $)	19,450	23,900	27,500	26,200
Quintile Shares (%)				
0 - 20	3.5	3.2	4.1	4.3
20 - 40	10.7	9.9	10.3	9.9
40 - 60	17.8	17.3	17.3	16.7
60 - 80	24.9	25.5	25.5	25.0
80 - 100	43.0	44.0	42.9	44.1
Gini Coefficient (%)	39.8	41.3	39.5	40.4
Population Shares by Range of Median Income				
< 50%	25.1	26.5	25.2	24.9
50 - 75	12.1	12.3	12.9	13.4
75 - 125	26.1	22.5	23.1	22.3
125 - 150	11.1	10.2	10.3	9.4
> 150%	25.6	28.6	28.6	30.0
75 - 150	37.2	32.7	33.4	31.7

TABLE 2: Inequality and Polarization Statistics, ELFP Individuals Age 15+, Labour Income

	Year			
	1967	1973	1981	1986
Mean income (1986 $)	16,950	20,150	20,700	20,400
Median Income (1986 $)	15,050	17,250	18,050	17,400
Quintile Shares (%)				
0 - 20	3.9	3.6	3.6	3.4
20 - 40	11.0	10.2	10.2	9.5
40 - 60	17.9	17.2	17.4	17.0
60 - 80	24.6	25.2	25.7	25.7
80 - 100	42.5	43.8	43.1	44.4
Gini Coefficient (%)	38.9	40.7	40.2	41.8
Population Shares by Range of Median Income				
< 50%	24.1	25.4	26.1	27.2
50 - 75	12.3	11.8	12.1	12.2
75 - 125	26.8	23.7	23.4	21.5
125 - 150	12.5	10.6	10.2	9.3
> 150%	24.4	28.5	28.1	29.8
75 - 150	39.3	34.3	33.6	30.8

Graph 1
Trends by Reporting Unit and Income Concept



a. Average Real Incomes ($000s)

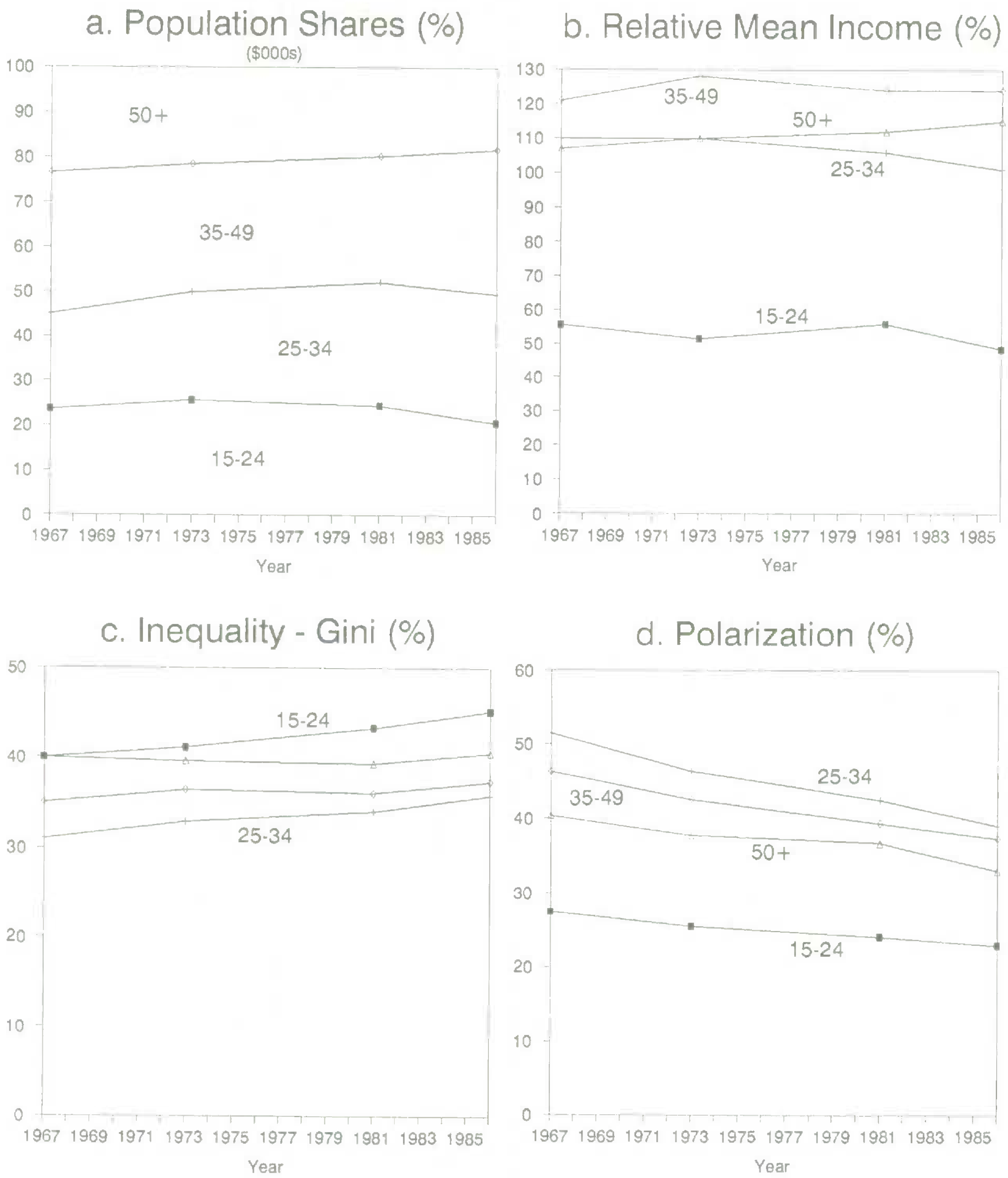
b. Average Real Incomes ($000s)

c. Inequality - Gini (%)

d. Polarization (%)

Graph 2 -- Trends by Age

ELFP Individuals Age 15+, Labour Income



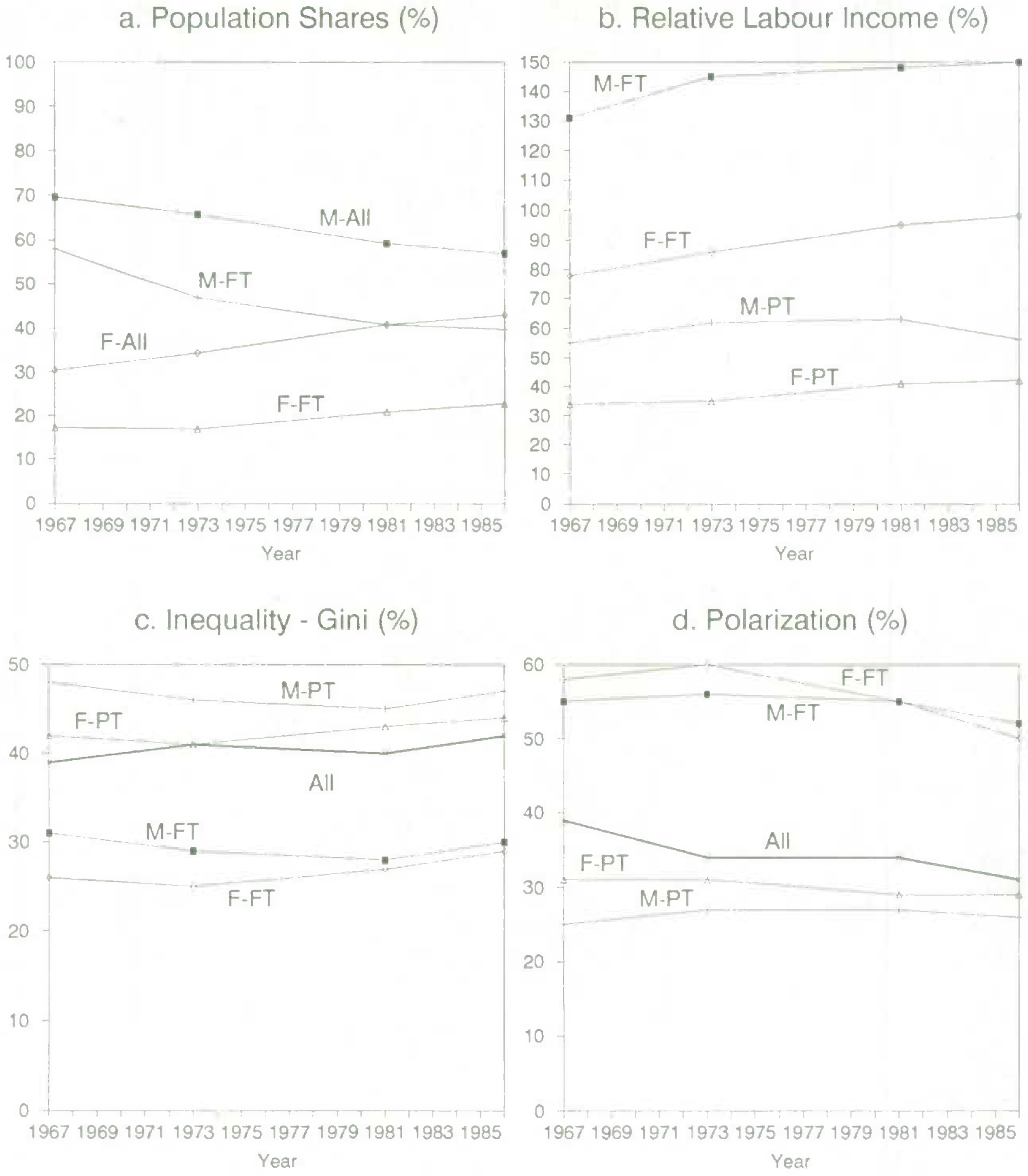
a. Population Shares (%)

b. Relative Mean Income (%)

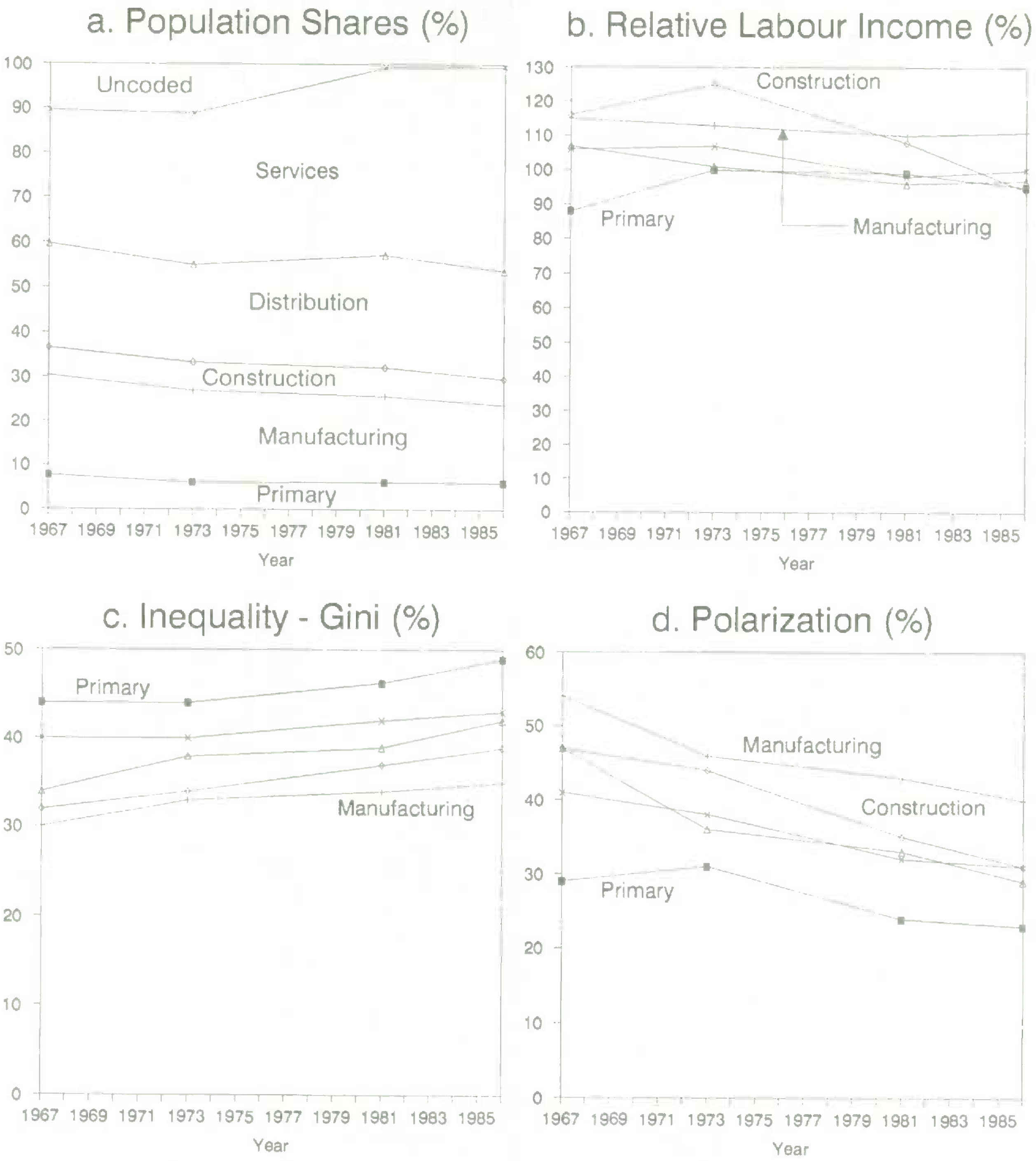
c. Inequality - Gini (%)

d. Polarization (%)

Graph 3 -- Trends by Sex aqnd Full-/Part-Time ELFP Individuals Age 15+, Labour Income



Graph 4 -- Trends by Industry
ELFP Individuals Age l5+, Labour Income



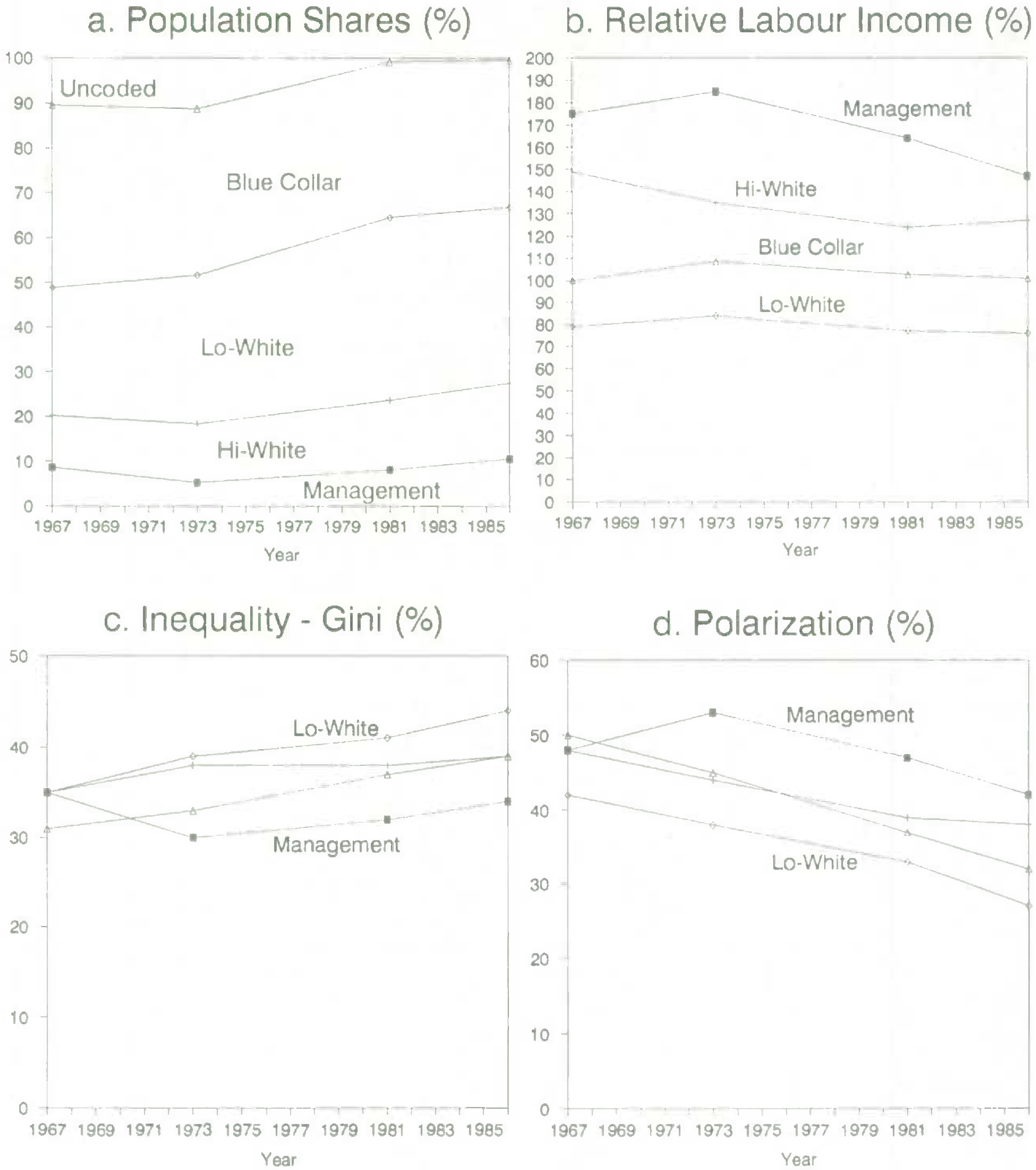
a. Population Shares (%)

b. Relative Labour Income (%)

c. Inequality - Gini (%)

d. Polarization (%)

Graph 5 -- Trends by Occupation
ELFP Individuals Age I5+, Labour Income



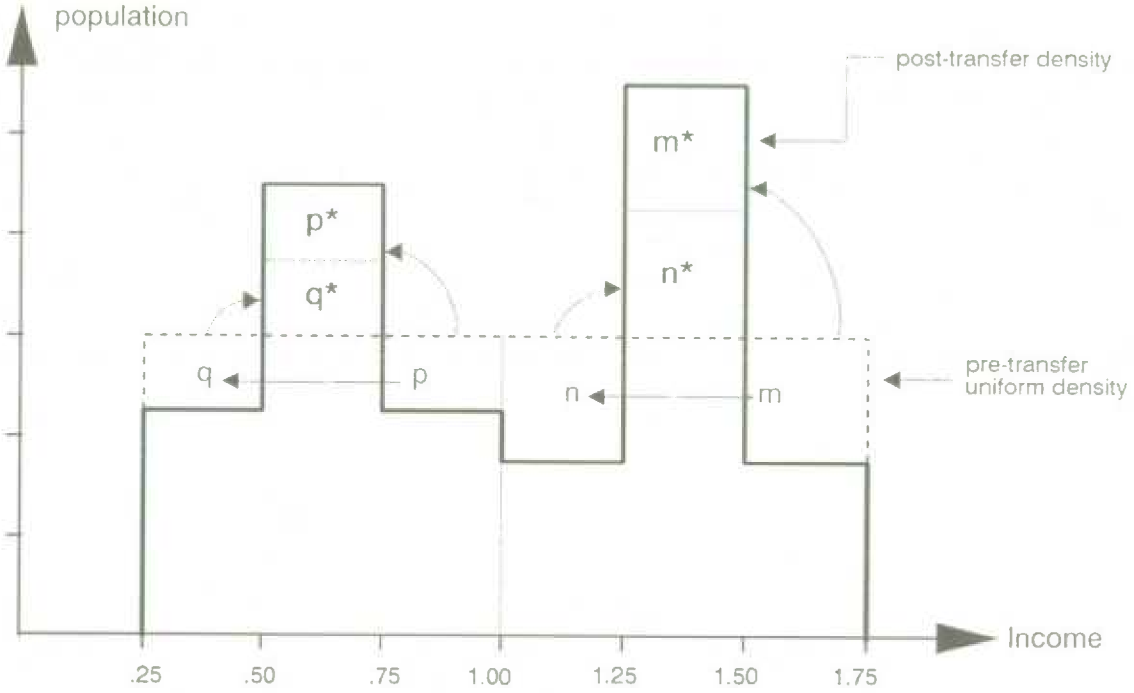
a. Population Shares (%)

b. Relative Labour Income (%)

c. Inequality - Gini (%)

d. Polarization (%)

Figure 1: Illustration of Polarization and Inequality



Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

DISCONTINUITIES IN TIME SERIES

C. Fortier[1]

SUMMARY

This paper provides an exhaustive list of the discontinuities that can affect time series of data, by type and degree of complexity. These discontinuities are numerous and must not be overlooked. The researcher must identify those that affect his particular series and must decide on the manner in which to proceed with his study (abandon it, make changes in specifications or make adjustments). The best tool to accomplish this is the sensitivity test which measures the impact of individual discontinuities observed. The extent of this impact is variable and an analysis of the development of seven series of measurements of the total fertility rate in Canada between 1950 and 1986 will permit us to make certain assumptions regarding their variation. In any event, it is important to place emphasis not on small differences in time but on major variations and sustained trends.

KEY WORDS: Discontinuities; sensitivity test; total fertility rate; adjustment of data; data analysis.

1. INTRODUCTION

Behind any analysis of time series, there is a researcher, an idea to be pursued and generally, one or more assumptions to be tested. The first step is to obtain long-term data that is comparable in time, pertinent and as close to the actual subject matter as possible. Unless the researcher is able to create his own data bank (by means of a retrospective survey, itself subject to discontinuities), he is faced with constraints and limitations of a technical nature. He is therefore limited to available data and must compile the results of various censuses, surveys and/or polls.

Data are always collected in order to provide answers to significant questions of current interest to users. However, data collection is subject to constraints such as the likelihood or ability of respondents to answer questions correctly, changes in public opinion, social and economic developments, and the political and financial conditions under which the respondent or organization is operating. The data collected must then be processed. Processing is dependent upon the questions selected on the infrastructure available, on the technical equipment and manpower, plus existing and available expertise. Finally, the findings are compiled and published, usually in the form of tables, by category (depending again on demand and requirements, as well as the classifications in use).

However, in time, concerns shift, requirements change and so does public opinion -- what was taboo yesterday can become fashionable today. The average level of education is improving with socioeconomic progress resulting in a higher standard of living and changes in lifestyle. Technological advancements and improvements in methodology are increasing the speed and quality of processing. Moreover, public awareness of new issues, such as matching of files or the confidentiality of data and respect for privacy, is posing new problems for data collection that cannot be overlooked. Finally, at the end of the line, publications are changing in number and content. This is obviously in response to collection and processing, user needs, diversification of interest and the growing refinement of classifications in use.

As a result, the data requested and published changes from one collection to another. The responsible agencies are aware of their requirements and of the concern for data that is comparable in time. As such, they must withdraw certain questions either temporarily or permanently when the questions fail to meet expectations and no longer provide any economic, social or political interest. Furthermore, they have to change the formulation of certain questions; this is a result of the evolution of social acceptability regarding the subject under consideration or in order to make them current. Finally, they must submit questions dealing with new themes on a regular basis, their total number depending on budgetary constraints.

Therefore, there is no certainty that, at a given date, the question asked will most adequately cover the problem identified, since it is the result of various compromises. Moreover, the degree of comparability in time of various questions dealing with the same theme is unknown since changes can occur at so many levels that it becomes difficult to evaluate the results in terms of quality and pertinence. The researcher is therefore faced with what we call "discontinuities in time series".

[1] C. Fortier, Demography Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

One more point remains to be made. The analysis of time series is not only based on data collected in the field: this type of collection is generally done only at intervals of a few years. It is customary, for the intervenin years, to resort to estimates produced by the researcher or borrowed from the work of colleagues or from official statistical publications. In the last case, the researcher may have to contend with changes in the estimation methodology or in the basic data used.

This brief overview is not very encouraging. Who would be brave enough to opt for a long-term analysis, aware that the series he is using are probably full of traps that can lead to error, either by themselves or in the computation of certain rates? How should the researcher proceed? First, he must identify the weaknesses in his data and try to correct them. To assist him in this task, we have attempted in this paper to draw an exhaustive list of possible causes of discontinuities, illustrating each case with a concrete example whenever possible. Since the number of discontinuities is large, we have classified them by type for the sake of clarity. Moreover, since they pose a problem for analysis, we have identified possible ways of dealing with them and, where required, adjusting them. Finally, we will attempt to measure the impact of various discontinuities on selected demographic indicators, by using a concrete example, and deduce, from the results observed, a scheme of operation of discontinuities in time series.

2. CAUSES OF DISCONTINUITIES

Discontinuities in time series operate at two levels between the reality we want to study and the observation by which we obtain our unfortunately imperfect data. Discontinuities at the first level are obvious and relate directly to the facts observed: they pose a problem at the outset. Discontinuities at the second level are more subtle, in the background, underlying, such that they are usually omitted when we speak of "discontinuities". They are behind the figures published, affecting their quality relative to the deviation in time between actual and observed values. First-level discontinuities therefore concern the format of the elements studied while second-level discontinuities relate more to trends.

Discontinuities at both levels affect time series in varying degrees of intensity and complexity depending, on the one hand, on the number of sources in time used and, on the other hand, on their simultaneous or non-simultaneous use. We will present the discontinuities in four parts, by function within the two previous criteria, and by increasing level of complexity: (1) changes in sources in time; (2) changes in sources in time chronologically; (3) changes in relationships between sources used jointly, in time; and (4) changes in sources used jointly, chronologically.

For each category, the discontinuities addressed will be divided by level. As they are numerous, we wish to apologize in advance for any possible omissions. They relate to many types of sources: surveys and polls; data registers; administrative files; censuses; studies, statistical computations by other authors and/or organizations. Some of these sources are part of a continuous registration program (data registers, for example vital statistics; certain administrative files such as the Family Allowance file of Health and Welfare Canada; provincial health care files; the immigration file of Employment and Immigration Canada) or a sporadic registration program (census; certain surveys such as the Canada Labour Force Survey). These sources, by themselves, can therefore serve as a base for chronological analysis. We will first examine the discontinuities that can affect sources in time.

2.1 Changes in Sources in Time

The constant concern for improving collection systems, whether they be governmental or private, necessarily leads to periodic gaps in time series, gaps that are either obvious (first-level discontinuities) or hidden (second-level discontinuities).

2.1.1 First-Level Discontinuities

The causes of first-level discontinuities are found at all stages of the process of data collection, preparation of questionnaires, processing and publication of data. They can even be found outside this process as we will see later.

The most obvious discontinuity is certainly that which results from a sudden discontinuation in the collection of data usually collected on a continuous basis, by reason of a war, famine, epidemic, revolution or, simply, budget cuts. It can also occur as a result of the loss or destruction of documents as was often found in historical studies. The gaps are all the more significant in the case of serious cyclical events when additional data sources are also missing and there is every reason to believe that these events affect the element or phenomenon studied, if not the target population. For example, the historical analysis of the number of patients in psychiatric institutions in France was interrupted between 1914 and 1919 because the First World War prevented the usual collection of data (Meslé and Vallin, 1981). These gaps, however, are usually of short duration as normal services are resumed as soon as the crisis is over.

The withdrawal of a question posed for a period of time has the identical effect although its origin is different. Such withdrawal can be temporary but is usually permanent as it is the result of a lack of social or economic interest in the concept. A recent example of this was the withdrawal from publications on vital statistics in 1978 of the table showing the proportion of hospital deliveries in Canada. There was no longer any reason for this table because the fundamental problem of maternal and infant mortality had become practically negligible and the proportions had been stable at nearly 100% during the last ten years. However, with the increase in home births, we may witness the reintroduction of this information in a few years. Another reason for withdrawing a question may be public pressure. The responsible agencies may want to ask the question and the researchers may want to know the answer, but the theme in question is irritating to the public. In the past, this has been the case with divorce, later with common-law unions and abortion. It is sometimes the case with ethnic or racial issues. In these cases, the question can be asked and then omitted several times according to the fluctuation of public opinion.

The modification of a question is slightly less obvious but just as problematic. Although data is available regularly in time, what it covers may vary. The discontinuity is more or less significant depending on the subject of the modification. It may be the concept behind a term or expression, such as the expression "legitimate birth order" in vital statistics in France. Between 1949 and 1964, it referred to "order by woman"; however, since 1965, it refers to "order in current marriage", which is totally different in the case of multiple marriages (J.-L. Rallu, 1986). The subject of the modification can also be the formulation of the question itself. Indeed, the desired information can be obtained in many ways and the question selected is but one of several possible formulations. Age is a good example. It can be obtained directly or through a question on the date of birth; these two options do not necessarily lead to the same answer as a number of studies have shown. Finally, certain selective questions that have no other purpose than to identify the respondent studied, but which can nevertheless be used in research, can be modified without too much concern in the belief that the change will be inconsequential. Yet in some cases, the change is fundamental. Usually, both the concept and the formulation are changed. A typical example of this is the person responsible for completing the census form in a household. This role, which used to be assigned to the "head of the household", i.e. the man in a couple, can now be assumed by a woman duly identified as "person 1" in Canada and as "reference person" in France.

Discontinuities can arise not only from questions but also from the target population which can be increased or decreased for various reasons. One reason is a political change in territorial limits (adjunction or cession of new territories for a country) or a simple statistical change (revision of boundaries of statistical regions). Another reason for changing the target population is simply one of definition. Broadly, the statistical agencies agree to propose two types of population: the "actual" population, i.e. that on site at the time of data collection, and the "de jure" population, i.e. that usually residing at the place surveyed. The change from one to the other in time can cause substantial deviations in data. Portugal is an example: "actual" prior to 1940 and "de jure" subsequently (Monnier, 1982). Other countries change the underlying concept while continuing to target the same type of population. This has been the case particularly in France which has seen an increase in its "de jure" population since 1962 by the inclusion of military personnel stationed outside the metropolis (Monnier, 1982). The definition of target populations by sources other than the census can also be modified, as evidenced by French statistics on the insane population which, since 1949, include only patients in "public institutions or private institutions serving as public institutions", excluding private hospitals which were previously covered (Meslé and Vallin, 1981).

Another cause of discontinuity in the long term is the frequency of collection, sporadic or continuous. Sporadic collection, in order to be comparable, must take place at regular intervals and therefore at fixed dates, which is not always the case. There are abundant examples of this as the dates of successive national censuses are often different (Monnier, 1982). As for continuous registrations, their totals must always correspond to the same time interval in terms of number of months and type of year (calendar, fiscal, census, school). In the Federal Republic of Germany, for example, the abortion count has covered a period of 12 months since 1977, but it covered only six months and a few days in 1976.

Once the basic data has been collected, it is processed and one of the processing steps is coding. Coding consists of assigning to the response of each individual a code that corresponds to a category in a classification established either by the responsible agency or the researcher himself or, most often, by international agencies specializing in the issue under consideration. The use of these international classifications is definitely advantageous as it permits a better comparison of national structures; it is widely recommended. However, whatever the classification selected, it runs the risk of being modified in time according to the evolution of society and knowledge. F. Meslé and J. Vallin (1981) encountered this problem in their study of causes of death in the long term in France (7th, 8th and 9th revisions of the WHO's International Classification of Diseases). The modifications may be slight or may affect the very principle of classification.

Classifications, groupings and modifications are not only done during the collection and processing stages. The characteristics selected in the tables further reduce the information that reaches the user unless he has the technical and, frequently, the financial capability to directly access the processed data. In short, the information already classified runs the risk of being grouped into wider categories for publication purposes, and these can vary from year to year. Finally, not all available data is published; some may be obtained simply on request. The choices made are obviously reconsidered with each publication, which again can lead to gaps.

Finally, one last discontinuity affects only sources of data from continuous registration. By definition, a data register is constantly modified by the addition and withdrawal of events affecting the target population or element, and this takes place, in theory, as soon as the events occur. Unfortunately, and this shortcoming is well-known for this type of source, the delays in registration are sometimes too long, forcing the updating of monthly files over several months. For example, the files of eligible recipients of family allowances for a given month are updated every six months for a period of two years (files M0023 and M0024). A chronological study based on such registers should use complete files or at least files in similar stages of completion (1st, 2nd, 3rd revisions) for the sake of comparability in update and quality. In this respect, this possible discontinuity could almost qualify as a second-level discontinuity.

2.1.2 Second-Level Discontinuities

Second-level discontinuities relate mainly to the quality of the information collected or available. Whatever the attention to detail during planning, data collection, data processing, whatever the experience and efficiency of the personnel involved or the scope of the means employed, it is almost impossible to obtain a perfect collection of data that satisfies all imaginable criteria of quality. And these are numerous. The total error and general quality of a particular source of data are respectively the sum of its various errors quite simply, and the difference between this sum and "1".

The first disparity between the actual element and the measured element results from an error in concept, i.e. the degree of pertinence of the questions asked. As shown in the previous section, many different questions can be used to determine the same state, each producing results with varying degrees of compatibility.

It is generally acknowledged that the greatest share of error is attributable to collection errors, whether they be a mismatch of the collection method and information sought, response errors, errors by the enumerator or errors of coverage.

Various data collection methods can be used to target the same subject of study in the same population. However, not all methods are well suited or likely to produce the same results. Take the case where we want to determine the total number of abortions undergone by Spanish women in a given year. Aware of the fact that abortion has been legal under certain conditions since 1985 and that each operation results in the completion of a statistical form, we could rely solely on the official figures. However, the abortion law has given rise to strong opposition in Spain and a large number of physicians are invoking the conscience clause to avoid the practice. As a result, the official number of abortions is very low (Monnier, 1986). However, Spanish women faced with these difficulties have been travelling to other more liberal countries to obtain abortions, such as the Netherlands and England (Monnier, 1982). The inclusion of abortions obtained elsewhere by Spanish residents in the official Spanish figures changes the picture completely. Moreover, clandestine operations are omitted!

There are three types of response errors: involuntary errors, voluntary errors and non-responses. The first type of error owes more to the passage of time, which causes forgetfulness, and to the respondent's interest in the subject matter than to bad faith on the part of the respondent. Also, the respondent is sometimes required to answer on behalf of other household members who he may not know very well. Voluntary errors are more subtle. They can result from the manner in which the respondent is approached, which can be more or less successful, from the purpose of the collection, from public opinion, which is sometimes negative regarding the information sought, or still, from the social, economic and political conditions existing at the time of the interview. An interesting example of the importance of this last point is the large proportion of the Albanian population of Yugoslavia that declared itself "Turkish" at the time of the 1953 Census in the hope of being able to emigrate to Turkey. Turkey only granted official residency to Turkish nationals (Islami, 1983). Non-responses, for their part, are the result of the respondent's inability to answer certain questions or refusal to answer. Misunderstanding of a subject or lack of opinion on a subject can lead to serious problems: people don't know what to answer, and the non-response rate tends to increase with the complexity of the question. Refusal to respond can be associated with one or more questions, or all the questions, and is attributable to lack of interest, the subject, existing socioeconomic conditions, the purpose of the collection, or the freedom of response allowed the respondent. These last two causes are well illustrated in the following example. The response rate of Jews in Tunisian censuses had not been very high. In 1941, the Tunisian government decided to tie the census to the status of Jews and, in order to ensure their participation, it threatened several sanctions against recalcitrants. The Jewish population enumerated was therefore much larger than usual and than the censuses of 1936 and 1946 indicated (Taieb, 1982). Response errors can also be attributed to errors by enumerators when they are used for collection.

Indeed, the enumerator, that intermediary between the respondent and the questionnaire and/or form to be completed, can introduce response errors either by omitting one or more questions or by forgetting to record or recording badly the response provided. The errors made by the enumerator, although they are involuntary, by his involvement in all stages of collection, combined with his lack of experience (as a general rule) or misunderstanding of basic concepts, account for most of the total error (Kalbach and McVey, 1971, p.11). In addition to these errors, the enumerator can be responsible for errors in classification, due to confusion over definitions, and errors in coverage, by forgetting targeted persons or including non-targeted persons. However, the latter errors, especially errors in coverage, are not eliminated by the non-use of enumerators or interviewers.

Errors in coverage are the difference between the gross undercoverage (proportion of targeted persons omitted) and the gross overcoverage (proportion of targeted persons counted twice or more, combined with the proportion of non-targeted persons enumerated). Their causes, other than attributable to enumerators, relate to the location and the season (it is difficult to enumerate in the Yukon in the winter!), the purpose of the collection, the respondents' reaction to this intrusion upon their privacy, the freedom of response allowed, and the reputation of the responsible person or agency. The entire target population may be affected or only a region or a certain category of people. The extent of the coverage error is not always known but, when it is, it can be and sometimes is adjusted. Discontinuities in time series can therefore derive just as much from variations in the degree of coverage in time, if there is no adjustment, as from the decision, at a given date, to adjust where it was not done previously, and from changes in the techniques used to measure the overcoverage and undercoverage rates or changes in the adjustment methods. The study of time series concerning the Australian or English population, for example, must take into account the fact that, for the past ten years or so, population estimates have been adjusted by the net undercoverage (gross undercoverage minus gross overcoverage).

Like collection errors, processing errors originate in the multiple procedures and operations: correction, adjustment and weighting. All three can cause discontinuities in the long term if their methodology is modified or if the subject of the procedure changes. The correction procedures concern only response errors. The procedure covers the decision to correct these errors and the method used to do so (smoothing of results, pro-rata or random allocation of responses among other forms with similar characteristics). Discontinuity can therefore result as much from the addition of a series of adjusted data to a non-adjusted series, as from a change in the method of adjustment in time. The adjustment procedures are extrinsic to the responses themselves and are applied in response to the concern for confidentiality of respondents and their government. For example, Statistics Canada has not released detailed cross-tabulations containing nearly empty cells since 1971. All final data is now adjusted by a "random rounding" process which hides smaller numbers behind a "0" or a "5". The weighting procedure concerns mainly surveys but can also be used with the census when part of the census is administered to only a sample of the total population. This procedure, which extrapolates to the entire population the results of the sample, generates what is commonly known as "sampling errors" or "random errors". The bias comes from the sample size and sampling method, the adjustment method used and its components (source of data, collection method...). Any change in one or more of these elements in time can affect the comparability of data. Before ending this paragraph on processing errors, we would like to point out the increasingly important role of technology in processing (data processing tools, mainframe computers, minicomputers or microcomputers). Their systematic use in all processing stages has greatly improved from year to year not only the management of operations but the quality of processing by the elimination of transfers and human error. Finally, it should be noted that random errors can result not only from weighting but also from the use of small numbers. It is a known fact that the smaller the size of the target population, the greater the random error. In a time series analysis in which the population grows in time, one should remember that the quality of data grows at the same time, especially where the beginning population is small and the end population is large.

The last cause of error we will mention here is related to the collection unit. Responsibility for collecting data in a given territory is sometimes divided among smaller geographic units, so that the partial results add up to the total results. Registration policy and procedures, as well as processing, can vary from one unit to the other, resulting in general inconsistencies.

Most of these errors affecting the quality of data cannot be measured or are not measured. Nevertheless, we know their importance and variability, a function not only of the very characteristics of the source or respondents, but also and mostly, of time. This point should not be overlooked in a time series analysis. Indeed, there seems to be a direct relationship between the data collection period and the quality of data: quality improves in time with the collection period. Without saying that old data is unusable, we are certain that the effort made, the infrastructure in place, the improvement of techniques, the growth of knowledge, financial means and especially experience cannot help but improve the quality of data. Moreover, these improvements have been supplemented by the implementation of more refined techniques for measuring the quality of information and identifying the nature of errors. In spite of this, and paradoxically, errors sometimes seem to grow with the improvement of techniques!

2.2 Changes in Sources in Time Chronologically

The use of a single source of data over a long period of time imposes numerous external constraints, as just demonstrated, all of which lead to discontinuities of varying degrees of significance. The decision to change the source can only compound the problem as the discontinuities of one source add to those of the other. It is true that the compounding of inconsistencies is not as simple as that: a modification can affect both sources at a same point in time and then, at least at that point, the change has no significant effect. A modification to the master source (the first one used), whatever the nature, can be so significant that it is better to resort to another source even if this entails numerous but minor discontinuities. Finally, sometimes one has no choice when the element studied is no longer available through the first source and was not available prior to the second source. In any event, it is important with respect to first-level as well as second-level discontinuities to determine the total variations generated by these discontinuities in the period studied.

2.3 Changes in Relationships Between Sources Used Jointly, in Time

It often happens that an analysis of some kind must use two or more data series jointly. This is the case, especially, when studying the evolution of rates, ratios and other social and economic indicators. In these cases, certain sources are used in the numerator, others in the denominator, thus lessening the impact of their respective discontinuities on the index observed. However, the joint use of data from various sources is not peculiar to these ratios. Such use is also made in the estimation of the population between two census dates, for example, and then, if the component method is used, it results from the addition of births and immigrants to and the subtraction of deaths and emigrants from the base population, generally that of the last available census. Then, the total discontinuity over the period is the total of the discontinuities of each data source, plus those relating them, for each point listed in A. This is also the case, of course, in the computation of ratios.

Discontinuities in relationships between sources can result from a change in the method relating them, but also from any change affecting one source without affecting the others or from any change altering the deviation between the sources without any appearance of change.

The concern for detail and accuracy has led researchers to continually refine their computation methods, at the same time altering the comparability from year to year. In Romania before 1975, for example, the life expectancy of the population was calculated for single years but subsequently, it was calculated for three-year periods (Monnier, 1982). However, it should be remembered that when a method, concept or orientation is modified by the statistical agency that disseminates the data, it is customary to perform retrospective adjustments or publish both series for a few years. This limits the impact of inconsistencies in the time series. There are many examples of changes that affect only one of the sources used jointly, such as the following: the computation of mortality rates in Canada based on the one hand on deaths (Vital Statistics, Statistics Canada), unchanged, and on the other hand, on the population estimated by a modified method (Demography Division, Statistics Canada). Finally, some changes may not affect any of the sources involved but only the deviation between them. This is much more evident when we examine concrete cases. The computation of fertility rates in Canada has always required the use of various target populations: the births in the numerator are derived from the actual population in the country and from Canadians abroad; the population in the denominator is of the "de jure" type. These concepts have not changed in any respect; however, the proportion of the actual population that is not resident has been growing steadily: refugees, students with visas, temporary work permit holders and, why not mention, illegal immigrants. The increasing deviation between sources can therefore falsify the findings of an analysis by distorting the trends.

Finally, we wish to point out that these jointly used sources become another source in themselves and that, in this regard, the caution stated in A relative to the use of a single source applies equally here. For example, an index can be affected by the modification of its target population; this was the case in Malta where the total fertility rate concerned all resident females before 1975 but later excluded foreigners.

We must emphasize the degree of completeness and type of data. Population estimates, for example, can be distinguished on three points. They can be postcensal, i.e. produced on the basis of the last available census and modifications to components since then, or intercensal, i.e. based on postcensal estimates and the preceding and following censuses. Their frequency can also vary: they are either done on a quarterly or annual basis. Finally, their degree of completeness also varies depending on whether they are provisional (3 to 4 months after the reference date), updated (8 months) or final (15 to 20 months). At the same time, the researcher must make sure that he is using final intercensal estimates at all times, or at least as much as possible. He must also seek a series of adjusted national indices, based no longer on postcensal estimates, as annual computations are, but on intercensal estimates or on the census depending on whether or not it is a census year.

2.4 Changes in Sources Used Jointly, Chronologically

A data series based on two or more sources can suffer in the long term from inconsistencies due to changes in any of the sources, in addition to the discontinuities mentioned above. Then, the possible discontinuities are those listed in B and C.

The discontinuities identified may vary in format and impact. Some can simulate an increase, others a decrease. The impact of discontinuities in a time series can be small, even if there are many discontinuities. On the other hand, a single discontinuity can have a major impact on numbers and especially trends. What can the researcher do? What tools are available to him to adjust these discontinuities and does he generally do this?

3. ADJUSTMENT OF DISCONTINUITIES

The reaction of demographers to time series already produced or to be produced, is not uniform. The spectrum of reaction is wide, ranging from consideration to non-consideration of discontinuities and including a variety of strategies.

It is not unusual to look through the literature in vain for comments on the quality of the data used and possible discontinuities in the long term. Of course, some studies are based on well-known data used over and over again. Unfortunately, this is not always the case. How, then, is the researcher to know if there are or were any discontinuities and, if so, how were they adjusted? The reader can form his own opinion based on the list of sources used. This solution is not necessarily bad: it all depends on the number, gravity and significance of the discontinuities.

It is therefore important, before starting his analysis, that the researcher determines the general quality of his series and its ability to measure the evolution of the element selected. Some researchers do this and, if they decide to keep their data as is, they do so in full knowledge of the facts. For example, Roussel (1983) noted a change in concept in an analysis of the evolution of households in the Netherlands since 1945. He measured the impact on the households involved and realized that it was insignificant (barely 2.5%). The impact was even less significant when compared with the large movements experienced by the households during the period. The author concluded that "the bias, while not negligible, did not invalidate the findings" (translation of Roussel, 1983, p.997). The decision not to adjust the data is therefore based on the measurement of the impact of the discontinuities. Others make this decision on the basis of the time at their disposal or the cost of adjustment. Nevertheless, they caution the reader against making hasty conclusions on the basis of small variations and list all the discontinuities. A typical example of this is the book published by Kalbach and McVey in 1971.

Among those who modify their initial work plan as a result of discontinuities, some limit the period of study, use broader classifications, abandon their work or make adjustments. Many researchers lacking time, funds or adequate techniques decide to limit detail in their analyses, in terms of length of period, characteristics or classification. J. Vallin (1983), for example, limited his description of the evolution of French mortality to the period 1950-1978, thus avoiding any possible distortions caused by the 5th and 9th revisions of the International Classification of Diseases. He nevertheless had to contend with distortions caused by the 6th, 7th and 8th revisions, which he attempted to limit as much as possible by using broader classification levels relatively common to the three series. Some researchers have no other option but to abandon the study considered. We cite, for example, the case of R.J. Lowe (1987) who had to give up his study of the evolution of full-time employment in New Zealand based on the 1981 and 1986 censuses, due to intercensal changes in the dividing line between full-time and part-time employment.

However, not all researchers have to abandon or change their initial idea. Many strive to link the data in time, to modify either the initial series or the auxiliary series, or at least to improve the quality of the information measured. There are no proven statistical methods to do this. There are only examples which can provide suggestions for researchers faced with identical problems. These numerous examples cover most of the discontinuities identified. We have chosen those that seem to be the most pertinent and innovative. The most complicated methods are not necessarily the best. Sometimes, the actual situation corresponds more to a simple form that is easy to approach. For example, F. Munoz-Perez and M. Tribalat (1984), in their study of the evolution of the number of mixed marriages in France between 1910 and 1982, simulated the events missing for the period 1932-1942 by a simple linear interpolation based on fragmentary data. Others have adjusted known figures in proportion to the missing population (Meslé and Vallin, 1981). Others have changed the scope of their analysis preferring a regional study that highlights inconsistencies (but only where they occur) to a general study that conceals them (Perrégaux, 1983). Still others have replaced numerous missing records by means of simple random selection among those available (Henry and Blayo, 1975). Another alternative is to refer to third sources which, although excluded from the main analysis, are useful for adjusting deviations and ensuring continuity in time. J.L. Rallu (1986), for example, dealt with a change in the concept of "birth order" during the period of his study by using the results of two family surveys conducted during the same period. Since these surveys provided data according to both the former and new definitions, Rallu was able to compute certain factors by period and five-year age group which were applied to the main data series. Other authors prefer to use an elaborate mathematical methodology. For example, Bergstrom and Lam (1989) advocate the use of cubic interpolations to solve the problem of inconsistency in the translation of age-specific data to year-of-birth-specific data.

These various examples point to a specific approach to the adjustment of incomplete time series. First, and this is an extremely important step, all decisions must be preceded by a specific evaluation of the situation, i.e. identification of discontinuities in the series at our disposal, their level and their impact on results. While identification of discontinuities is not too difficult, it is a different matter with measuring their level. As a general rule, the researcher has no control over the data base he did not produce. The general quality of published data is only known to him thanks to studies made by the responsible agency. Of course, he can conduct some comparative studies himself but, for certain variables (undercoverage for example), the researcher is dependent upon the decisions of others. Then, the results provided may not meet his expectations and, as is often the case, are unusable because they are too superficial. However, this problem is not insurmountable. Some of the examples cited show that, sometimes, it is not necessary to adjust the data when the impact of the discontinuities on the results is negligible. Then, the impact on individual cases remains to be measured.

4. IMPACT OF DISCONTINUITIES

The impact of discontinuities can be measured quite simply by the use of sensitivity tests. This involves applying to the data a series of assumptions relative to their error rate, adjusting them to take these into account and observing the impact of these adjustments on the results. When the impact is negligible, all is well. Otherwise, the solution is to adjust the data if the actual error rate is known. The final decision to continue the study rests with the researcher: he alone is the judge. This decision must be made in full knowledge of the facts and if the researcher intends to carry on, he should caution the reader against hasty conclusions due to small variations in the series and should explain their probable source.

However, we should not be pessimistic. We will conclude this paper by attempting an evaluation of the impact of certain types of discontinuities on various demographic indices. In order to do this, we will use the concrete example of the evolution of the total fertility rate (TFR) in Canada between 1950 and 1986. This index is obtained annually by summing the age-specific female fertility rates without distinction of marital status. Fertility rates regardless of marital status relate births to women at age x to the total number of women at the same age.

There are many options available to us regarding the data to be used. Total fertility rates are computed on a regular basis and appear in various official publications. At the international level, the Demographic Yearbook of the United Nations provides fertility rates without distinction of marital status by five-year age groups for Canada for the period chosen; the index sought is obtained by multiplying these rates by 5 and summing them. These rates exclude births from the province of Newfoundland but are based on the total Canadian population. Canadian citizens temporarily residing in the United States are included (U.N. 1979; 1987). At the national level, Statistics Canada (and previously the Dominion Bureau of Statistics) publishes vital statistics annually as well as certain simple indices derived from them, including the total fertility rate. Two series are available up to 1972: the indices as published annually and based on post-censal estimates of the population; and the indices revised after each census and thus based on intercensal estimates of the population. This last series has not been produced since 1972. We could also decide to make our own calculations of these indices based on final birth data and revised intercensal population estimates. Since Newfoundland has no data on birth by age of mother, it is excluded from the calculation of the previous Canadian indices. We could estimate these births and recalculate the TFR including that province. Certain sensitivity tests (simulations) are also possible. Given that the target populations for recording vital statistics and for the census are different (vital statistics = population usually residing in Canada + certain Canadians abroad, including diplomats and the military; census = "de jure" population), we could decide to estimate the deviation between them, overestimate the number of women exposed to the risk of reproduction and recalculate the rates. Finally, since undercoverage is a known characteristic of Canadian censuses, we could adjust the actual population previously increased, based on coverage rates obtained by the Reverse Record Check (although these rates are not very reliable because the margin of error is so large).

Thus, we could have seven different series of data for the same analysis. Their comparative analysis would permit us to measure or at least evaluate the impact of their differences (which, in fact, are discontinuities in the series) on the index studied and its evolution in time. These discontinuities are:

1. Change in the method of calculating the index: The United Nations uses rates by five-year age groups; the other series use rates by single years of age.
2. Change in the method of calculating intercensal populations: estimates revised after each census for those used by Statistics Canada; estimates revised after the fact, by using the same methodology for the entire period.
3. Change in the type of population estimated: postcensal or intercensal population based on the two series published by Statistics Canada.
4. Change in the target population in the denominator: usually "de jure" population but, for the last two series presented, the population targeted by vital statistics.
5. Change in the territory: inclusion or exclusion of Newfoundland.
6. Change in the quality of population data: with or without adjustment for undercoverage.

It should be noted that these six discontinuities are not the only ones that affect the data: recording of vital statistics is a provincial matter while the census and resulting estimates fall under federal jurisdiction; the definition of "live birth" was changed in 1959 and the new definition was applied to the provinces at various dates; finally, since 1981, the census of Canada has no longer been dependent on the work of enumerators as each household now completes the questionnaire and returns it by mail. However, we will limit our study to the impact of the six discontinuities mentioned previously because the impact of those excluded on our series is the same or is much more complicated to measure.

Graph I shows that, whatever the series, the trend of the total fertility rate between 1950 and 1986 is the same: steady growth up to 1957 followed by a sharp decline up to 1973, followed by a much slower decline. However, there are some differences in the United Nations series adjusted for undercoverage. A detailed examination of the annual trends sometimes shows some variations. In 1983-84, all the series show a slight increase except for the series based on the actual de jure population which shows a levelling off. The same trend is observed in 1974-75.

The impact of the six discontinuities mentioned previously can be highlighted by comparing pairs of series with a single varying element. The impaet of the TFR computation method seems to be significant: the transition from rates by single years of age to rates by five-year age groups greatly reduces the magnitude of this index (approximately 2.5%), and this applies over the entire period. The comparison of the two series based on intercensal estimates (whether revised or not) shows no deviation or only a very slight deviation. The change in the computation method therefore has very little impact. This is not surprising sinee, by definition, and regardless of the method used, intereensal estimates are derived from the preceding and following eensuses which do not vary. The impaet of the use of posteensal or intercensal series is also negligible. Here again, this is not surprising since it has been shown that the difference in Canada between the two types of population is generally less than 1% (Statistics Canada, 1987). The ehange in target population from "de jure" to "actual de jure" is slightly more signifieant, especially for the last few years. This is attributable to our adjustment assumptions (based on our fragmentary knowledge): the number of temporary Canadian residents for more than one year (refugee claimants, students, work permit holders, diplomats, visitors) increases continually. Finally, the simulated adjustment of the undercoverage of the various population eensuses has a major impact, espeeially between 1950 and 1965 and since 1976. This adjustment is based solely on known undereoverage data published in relevant doeuments. The method of obtaining this data varied before 1961 but, since then, the method used has been the Reverse Record Check, an operation that seeks, among the population enumerated, a sample of people that should have been enumerated. The size of the sample has increased eonstantly sinee then but the error rates have remained high. We wish to repeat that this is only a simulation. The impact between 1950 and 1965 is signifieant because it affects the index by an average of 4.4%. The high rate of undereoverage of the population 20-24 years of age therefore has an impact here, as does that of the 1961 census (3.3%). The impact sinee 1976 is slightly smaller (3.5%), even if it is inereasing. This difference between the two periods is not due to a decrease in the undercoverage rate in time but rather to an increase in the mean age of mothers at birth (women aged 25-29 are undercounted less than women aged 20-24). The index revised to inelude Newfoundland is absent from the graph because it was computed only for two years. The adjustment, however, has no impact because the indices are identical to those based on revised intercensal estimates. It is true that the population of Newfoundland is small in relation to the population of Canada as a whole.

Therefore, with a few exceptions, the impact of the discontinuities on indices such as the total fertility rate is rather small. Taking a large number of different rates into aceount in this type of index in a way dilutes any error that might be associated with one of them in particular. What impact would certain discontinuities have on speeific rates, for example at the age where they are the most pronounced? Sinee undercoverage has a significant impact on the TFR and the impact is particularly high in the 20-24 age group, we will study the evolution of the fertility rate without distinction of marital status at age 22 between 1950 and 1987. We will limit our comments to four series as the three official series published do not provide any information on rates by single years of age.

Graph II shows the eomparative evolution of the series based on the revised intercensal estimates, the aetual de jure population, the adjusted undercoverage and the addition of the province of Newfoundland. As in the case of the TFR, the first two series are almost identical, even if the second series is slightly lower at the end of the period. In fact, the deviation between them is smaller than for the TFR due partly to the fact that the median age of the female population added to the actual de jure population is 27.7 years and that the maximum impact on the rates is not at age 22. However, as expected, the impact of the adjusted undercoverage is more significant than for the TFR, the mean deviation between 1977 and 1986 being 4.35%. As for the impact of the inclusion of the province of Newfoundland, it is small but noticeable, with the highest rates observed in 1980 and 1981, the only years available.

The age-specific fertility rate, like the TFR, is a ratio between two sources. We will now measure the impact of certain discontinuities on the female population aged 22. The previous series, except that including Newfoundland, will be compared. Graph III illustrates the trends. Again, the trends are similar but the position of the curves is the inverse of that shown previously. This is not surprising since the population serves as the denominator in the TFR and the age-specific rate, thus the inversion. The evolution of the intereensal population estimates follows that of the actual de jure population at least until 1976, after which the latter increases more rapidly. In 1986, the gap between the two series is a respectable 1.6%. The impact of the adjusted undercoverage is interesting. Not only is a widening gap becoming apparent at this point between the other two series and the adjusted undercoverage series, but while the others seem to show a decline in the female population aged 22 since 1983, the adjusted series suggests a levelling-off of this population. In 1986, the deviation between intercensal estimates and the population adjusted for undercoverage was 5.8%.

To summarize the previous comments, it is useful to combine in a single table the observations on the impact of the various discontinuities considered (Table 1). Even though the table concerns only specific demographic measures and seleeted discontinuities, it is very informative. First, it shows that the stronger and more distant the aggregation from the source of discontinuity, the weaker the impact. For example, the impact of a given discontinuity is much larger on the element directly affected, in this case the population, than on any index using this population. Moreover, the impact is reduced if the index eomputed does not solely use the population of the category with the high error rate. The extent of the impaet therefore depends on the degree of aggregation of the element studied in relation to the point where the discontinuity begins. However, the impact of this discontinuity on the element measured is also important. We can state without great risk of error that the impaet of undercoverage is certainly greater on the TFR than on the life expectancy at birth because birth, Impact of Selected Types of Discontinuities on Three Demographic Indicators

Discontinuity	Deviation Measured (%)		
	TRF	Fertility Rate at age 22	Female Population at age 22
. method of calculating the index	2.5	n.a.	n.a.
. method of calculating intercensal estimates	negligible	n.a.	n.a.
. type of population: postcensal or intercensal	negligible	n.a.	n.a.
. target population: de jure or actual de jure	1.2	1.0	1.6
. territory	negligible	1.7	n.a.
. census undercoverage	4.0	4.4	5.8

n.a. = not available

as opposed to death, occurs at an age where undercoverage is high. Table 1 also reminds us that some discontinuities have little or no impact and that individual impact is intrinsically tied to the very characteristics of the discontinuity. Here, the modification of the territory has little impact because the population of Newfoundland is small in relation to the population of Canada as a whole. What impact would the exclusion of the province of Ontario have caused?

The example given here concerns indices that can only increase or decrease without being negative: the format alone can vary, not the direction. The phenomenon of migration is different: the impact of discontinuities can even modify the direction of the net total movement. The study of such phenomena therefore requires additional precautions.

It would have been desirable to draw an exhaustive list of first-level discontinuities and indicate their impact on all possible demographic indices, with supporting figures and for different situations. Of course, it was impossible for us to do this but it should be done directly by the researcher at the beginning of his work.

5. CONCLUSION

Any time series has the potential for numerous discontinuities. These range from discontinuation of publication of a variable to modification of quality of data. Reaction to discontinuities also varies among researchers: abandonment, limitation of study and adjustment of incomplete information.

In fact, there is no magical recipe. The researcher is the sole judge in matters of discontinuities. Each study is particular in the sense that, not only is the investigated theme important in the final decision, but so is the environment: available sources, human and even financial resources, contacts and the researcher's position. Researchers who have access to basic data (contained in forms or questionnaires) are not in the same position as those who can only obtain costly detailed tables, even when working on the same subject.

The study of the development of the total fertility rate conducted as an example permitted us to draw certain conclusions which, although incomplete, illustrate well that the impact of discontinuities is not always great. It varies depending on the rate of error, the type of discontinuity, the element studied and its degree of aggregation, and the size of the population affected by the discontinuity in relation to the total target population. Important points to remember: geographic level and, especially, size of the population studied.

In any event, it is important to change our view on time series of data, to use caution and pay particular attention to the methods used. Finally, we should not attach too much importance to small differences in time, but rather we should place emphasis on major variations and sustained trends.

ACKNOWLEDGEMENTS

The author wishes to thank Mr. Jean Dumas and Mrs. Margaret Michalowski and Jeannine Perreault for their valuable advice and suggestions. The technical assistance of Mrs. Monique Charbonneau, Denise Laflamme and Lili Stevenson Roy is also greatly appreciated.

REFERENCES

Bergstrom, T., and Lam, D. (1989), "Recovering event histories by cubic spline interpolation", *Mathematical population studies*, 1, 327-355.

Henry, L., and Blayo, Y. (1975), "La population de la France de 1740 à 1860", *Population*, 30, 71-122.

Islami, H. (1983), "La population albanaise de Yougoslavie: accroissement numérique et répartition spatiale", *Population*, 38, 166-173.

Kalbach, W.E., and McVey, W.W. (1971), *The demographic bases of Canadian society*, Toronto: McGraw-Hill Ryerson Limited.

Lowe, R.J. (1987), "Comparing 1981 and 1986 Census Labor Force employment and unemployment data", *New Zealand population review*, 13, 27-34.

Meslé, F., and Vallin, J. (1981), "La population des établissements psychiatriques: évolution de la morbidité ou changement de stratégie médicale", *Population*, 36, 1035-1068.

Monnier, A. (1982), "La conjoncture démographique: l'Europe et les pays développés d'outre-mer", *Population*, 37, 911-940.

____ (1986), "La conjoncture démographique: l'Europe et les pays développés d'outre-mer", *Population*, 41, 823-845.

Munoz-Perez, F., and Tribalat M. (1984), "Mariages d'étrangers et mariages mixtes en France. Évolution depuis la première guerre", *Population*, 39, 427-462.

Perrégaux, J.-C. (1983), "La hausse de la mortalité infantile en U.R.S.S.: mythe ou réalité", *Population*, 38, 1050-1055.

Rallu, J.-L. (1986) "Descendance des générations françaises et probabilités d'agrandissement", *Population*, 41, 763-802.

Roussel, L. (1983), "Les ménages d'une personne: l'évolution récente", *Population*, 38, 995-1015.

Statistics Canada (1987), *Population Estimation Methods*, Canada Catalogue, 91-528E Ottawa: Minister of Supply and Services Canada.

____ (annual), *Vital Statistics*, Ottawa.

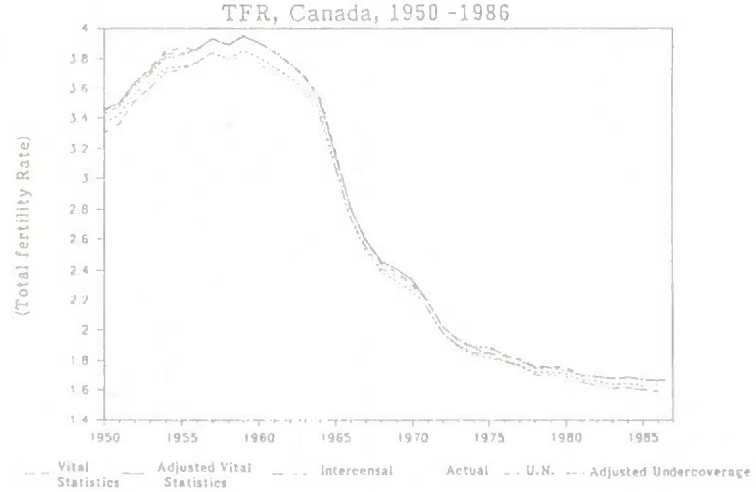
Taieb, J. (1982), "Évolution et comportement démographique des Juifs de Tunisie sous le protectorat français (1881-1956)", *Population*, 37, 952-958.

United Nations (1979), *Demographic Yearbook 1978, Historical Supplement*, New York: United Nations.

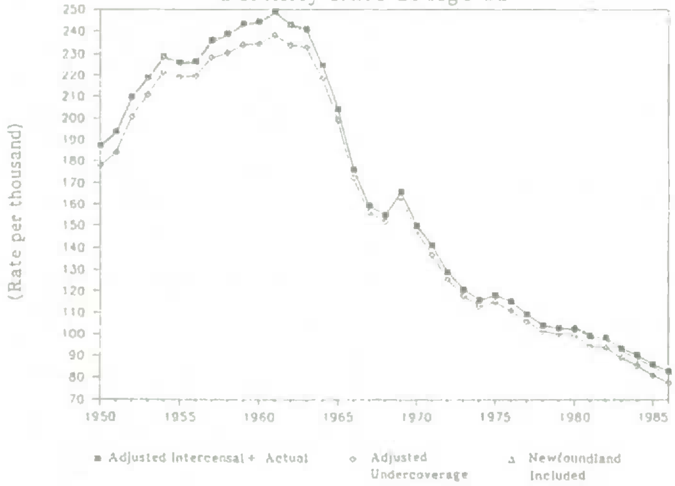
United Nations (1987), *Demographic Yearbook 1986, Special Topic: Fertility Statistics*, New York: United Nations.

Vallin, J. (1983), "Tendances récentes de la mortalité française", *Population*, 38, 77-105.

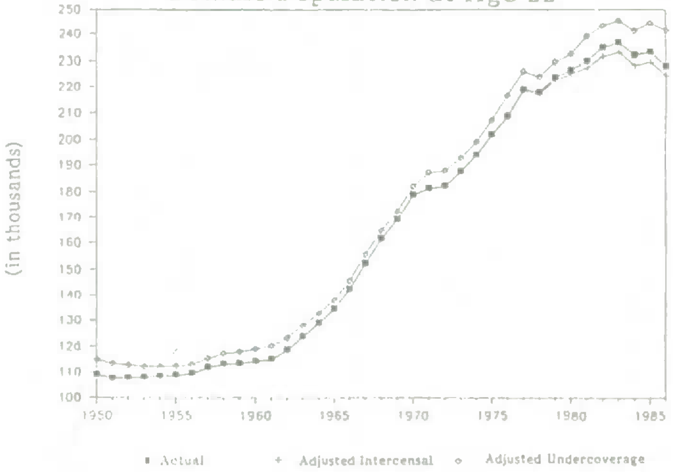
GRAPH I
TFR, Canada, 1950 -1986



GRAPH II
Fertility Rate at Age 22



GRAPH III
Female Population at Age 22



PART 7

ECONOMETRICS

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

LONGITUDINAL ECONOMIC DATA AT THE CENSUS BUREAU: A NEW DATA BASE YIELDS FRESH INSIGHTS ON SOME OLD ISSUES

Robert H. McGuckin*

ABSTRACT

This paper has two goals. First, it illustrates the importance of panel data with examples taken from research in progress using the U.S. Census Bureau's Longitudinal Research Database (LRD). Although the LRD is not the result of a "true" longitudinal survey, it provides both balanced and unbalanced panel data sets for establishments, firms, and lines of business. The second goal is to integrate the results of recent research with the LRD and to draw conclusions about the importance of longitudinal microdata for econometric research and time series analysis. The advantages of panel data arise from both the micro and time series aspects of the observations. This also leads us to consider why panel data are necessary to understand and interpret the time series behavior of aggregate statistics produced in cross-section establishment surveys and censuses. We find that typical homogeneity assumptions are likely to be inappropriate in a wide variety of applications. In particular, the industry in which an establishment is located, the ownership of the establishment, and the existence of the establishment (births and deaths) are endogenous variables that cannot simply be taken as time invariant fixed effects in econometric modeling.

KEY WORDS: Longitudinal; Panel data; LRD; Microdata.

1. INTRODUCTION

"You can't always get what you want, but if you try sometime ... you get what you need." (Let It Bleed, 1969, Mick Jagger and Keith Richards)

This paper has two goals. First, it illustrates the importance of panel data with examples taken from research in progress using the U.S. Census Bureau's Longitudinal Research Database (LRD). A panel data set is one that contains multiple observations on economic entities over time. For example, an establishment panel data set might have observations on shipments across individual plants linked over time. In contrast, time series data usually refer to observations over time on an aggregate economic variable, such as total industry shipments or U.S. national income. The advantages of panel data arise from both the micro and time series aspects of the observations.

Although the LRD is not the result of a "true" longitudinal survey, it provides both balanced and unbalanced panel data sets for establishments, firms, and lines of business.[1] The LRD enables researchers to conduct many essential studies heretofore considered impossible. In this sense, "you get what you need."

The second goal is to integrate the results of recent research with the LRD and to draw conclusions about the importance of longitudinal microdata for econometric research and time series analysis. The discussion focuses on research involving the behavior of firms and establishments. This also leads us to consider why panel data are necessary to understand and interpret the time series behavior of aggregate statistics produced in cross-section establishment surveys and censuses.

Most economic modeling is based on theories concerning the behavior of individual economic agents. Estimation and inference based on aggregate data involve assumptions about the homogeneity of the individual entities making up the aggregate. For example, a typical assumption might be that the distribution of the entities with respect to a particular variable such as efficiency or industry classification remains constant over time. This study indicates that typical homogeneity assumptions are likely to be inappropriate in a wide variety of applications. The evidence illustrates a basic point: The industry in which an establishment is located, the ownership of the establishment, and the existence of the establishment (births and deaths) are endogenous variables which cannot simply be taken as time invariant fixed effects in econometric modeling.[2]

The importance of panel data sets for economic research cannot be overestimated. Many economic issues simply cannot be addressed in the absence of panel data. As noted, these issues include a wide range of questions involving behavior before and after particular policy actions or other changes in the circumstances or environment of economic agents. Panel data sets also provide a unique vehicle for calculating microlevel measures of gross changes that are often missed in the aggregate statistics.[3]

* Robert H. McGuckin, Chief, Center for Economic Studies, (CES) Bureau of the Census, Washington, D.C. 20233. This paper does not necessarily reflect the opinions of the Census Bureau. The author acknowledges the comments and many useful discussions with Timothy Dunne and John Haltiwanger. Helpful comments from Stephen Andrews, Scott Schuh, James Monahan, Sang Nguyen, and Mary Streitwieser are appreciated. Stephen Andrews and Robert Bechtold did the data preparation and Janice Bryant and Terry Norris provided excellent typing services.

New evidence from CES research suggests that measures of gross change are important for many issues which have generally been examined with data on net changes. For example, some new work on job turnover finds that gross job reallocations are important in both a time- series (business cycle) and cross-section (across establishments and industries) sense. Work dealing with the entry and exit of firms and plants reaches similar conclusions. Analysis of one measure of turnover in industrial markets is used to contrast the importance of gross with net flow measures commonly available for analysis.

While the mechanisms at work are not completely understood, there are several reasons for expecting gross change measures to have important economic impacts. First, change typically requires resources and therefore measures of gross change provide a basis for measuring and understanding such costs. Second, the evidence that ownership change affects performance suggests that gross turnover measures provide important information on competitiveness.

A third reason for examining gross changes is that they provide a basis for determining if aggregate movements are being generated by a large or small segment of economic entities. Knowing how broadly based are the forces behind aggregate movements is important for policy makers. Longitudinal panels are required to address the issues involved in each of these examples.

Because the LRD is relatively new, a brief description adds some concreteness to the discussion. It also provides a basis for evaluating the LRD as a source of panel data.

2. THE LRD

The LRD is constructed by linking together individual establishment records from the Census of Manufactures (CM), which takes place every 5 years, and the Annual Survey of Manufactures (ASM), conducted each year. At present, the LRD has substantially more than 2 million manufacturing establishment-year records including information on over 800,000 different establishments in the 1963-86 period. When the 1987 census is included in the database, the number of unique establishments will likely jump to over 1 million.

Table 1 provides a tabulation of the number of establishments in the LRD in each year. Each census year, 1963, 67, 72, 77, and 82, contains well over 300,000 establishments of which about two thirds are actually surveyed. The administrative record cases, those which are not directly surveyed, represent small establishments (primarily establishments with less than 5 employees) which have little impact on aggregate industry totals. In non- census years the LRD contains roughly 70,000 establishments in the period 1973-78 and 55,000 after 1979 when there was a major redesign of the ASM.

The probability that any plant is sampled for the ASM is directly related to its size. However, the relationship is complicated. Large establishments, those with more than 250 employees, are sampled with certainty. Among the remaining smaller establishments (those with employment less than 250 and greater than 10), establishments are sampled with probabilities directly related to employment size except that there is an attempt to exclude those establishments sampled in one panel in the following panel.[4] This rotating panel design is to reduce the reporting burden on small plants. New panels are chosen every 5 years with the primary aim of obtaining accurate estimates of aggregate industry variables such as shipments.[5]

2.1 Cross-Section Design

The LRD data are collected from surveys and censuses that are cross-sectional in design and processing. While the processing procedures include previous year values in the edit sequences, there are few time based edits. As an example of the cross- section design, some large establishment reports are split into two or more establishments when the establishment produces a variety of distinct outputs. This procedure increases the precision of industry aggregates in the cross-section, but reduces the accuracy of establishment linkages across time by making it more difficult to trace individual plants.

The rotating panel design and the fact that most establishments are not sampled for the ASM do not in principle have any effect on the cross-section aggregate estimates. However, it does make following establishments over time more difficult and reduces the number of establishments who have continuous data for every year. The effects of this design on the availability of consistent yearly panels from the LRD are significant. For the 1972-86 period there are only a little more than 16,000 establishments that have data in every year in the LRD, less than 5 percent of the establishments in existence in any year.

Establishments not sampled in the ASMs appear only in the CMs. With 5 censuses available and another (1987) to be available soon, the possibilities for research based on balanced panels with observations at 5-year intervals are good.[7] Information on the composition of the linkages available for the census years 1972, 77, and 82 is presented in Roberts and Monahan (1986). They show that of the roughly 600,000 unique establishment records identified in these 3 years, approximately 133,000, or 22 percent, are present in all 3 years. These data were extended to the 1963-82 period by Dunne and Roberts (1986). For the 1963-82 period approximately 66,000 linked establishments are available to form a balanced panel. Although attrition will reduce the panel number by 1987, there still should be over 50,000 establishments observed continuously from 1963-87.

2.2 Data

The LRD contains a variety of information on individual establishments. Most of the data are reported on a yearly basis, but employment and hours worked are provided quarterly. By and large, the data contained in the LRD relate to production and various classification and identification characteristics of establishments. The latter category includes information on the plant's ownership, location, age (for some plants), product and industry structure, and various status codes which identify, among other things, birth, death and ownership changes. These identifying codes are used in developing both the longitudinal plant linkages and ownership linkages among plants.

Most of the data collected for each plant provides information on the inputs or outputs of the plant. A detailed description of the individual data items would be too lengthy to include here, but can be found in the LRD Technical Documentation available from CES that maintains and updates the LRD. However, the list of variables shown in Table 2 gives a good idea of the breadth of coverage. On the input side the LRD contains data on major factors of production; labor (production and other), capital, materials, and purchased services.

The output data include value of shipments reported for each 7-digit product in census years and at the 5-digit level of detail in ASM years. Related information, such as value added, miscellaneous receipts, inventories, value of resales, and receipts for contract work are also available for each establishment.

For the most part price data can be derived in census years in the form of unit values.[8] Outside of census years the quantity data to calculate unit values is not available in the LRD. This means that price series for purposes of, for example, deflation in production function estimation must be based on industry level price series. Such a series is published by the U.S. Department of Commerce based on Bureau of Labor Statistics (BLS) data. This series has been used by several researchers for purposes of deflation.[9]

3. LRD RESEARCH AND TIME

The research program at CES emphasizes projects that exploit the longitudinal characteristics of plants and firms. Many projects are measurement orientated. They establish important sets of "stylized facts" that form the basis for more substantive hypothesis testing. Examples of work in this category are studies by Dunne, Roberts, and Samuelson (1988, 89b) dealing with patterns of firm entry and exit and gross employment flows, respectively. Both of these studies used 5 year panels formed from census year data in the LRD. Other work in this category is reported in Davis and Haltiwanger (1989), where new measures of gross and net employment fluctuations at yearly intervals are constructed. Other studies at CES are oriented toward testing particular hypotheses. In this category of work there are various studies examining the importance of ownership changes on plant and firm performance that exploit the longitudinal structure of the LRD. Examples in this category are work by McGuckin and Andrews (1988), and Lichtenberg and Siegel (1988, 89a, 89b).

In what follows no attempt is made to be exhaustive in describing LRD research. For example, a wide variety of work on productivity measurement is not discussed in any detail. (Several studies have been published in the last 2 years and there are several other major projects underway.) The purpose is to illustrate the types of research for which panel data such as those contained in the LRD are essential. But, even more important, is the evidence that these types of analysis are crucial to understanding important economic phenomena and making informed policy judgements.

3.1 The Behavior of a Plant Over Time

It is useful to begin with a simple model to characterize the performance or behavior of an economic entity such as a plant or firm. For concreteness, assume that the i_{th} plant's performance at time t, Y_{it}, can be described by the relationship

$$Y_{it} = \alpha + \mu_j + \lambda_t + \Sigma_s \beta_s X_{sit} + \varepsilon_{it}$$

where X_{sit} are exogenous explanatory variables, α represents plant level fixed effect common to all plants, μ_j is a time invariant fixed effect such as ownership, industry, or location which is common to a group of plants, λ_t is a time varying fixed effect that is constant over individual plants, and ε_{it} is an error term. This simple model of plant performance can be used to characterize the issues of interest.

One important question is what can be controlled with the fixed effects specification. Of particular interest here is the question of what are the time invariant effects which can be represented by μ_j. One obvious candidate is the industry classification of the plant. Another is the ownership of the plant. Neither candidate is satisfactory.

3.1.1 Ownership Changes

The size and scope of the recent merger movement makes clear that plant ownership is often changed. Treating ownership characteristics as time-invariant is appropriate if the plant's behavior remains relatively unchanged before and after the ownership change. But, studies with the LRD indicate that ownership changes have dramatic effects on operating performance, whether measured at the plant or firm level.

McGuckin and Andrews (1988), find that line of business market shares increase relative to those of lines of business not experiencing a merger, particularly for complete firm takeovers. Lichtenberg and Siegel (1988, 89b) find improved plant productivity following ownership change. Furthermore, they are able to associate most of this gain with fewer administrative employees and lower wages for them following ownership changes

(Lichtenberg and Siegel (1989a)). These kinds of "event" studies are impossible without a panel of observations on individual establishments or firms.[10]

3.1.2 Primary Industry Affiliation

Various work has also shown that the industry category of establishments changes frequently. Approximately one third of the panel of more than 16,000 establishments continuously observed from 1972-86 experienced a switch in their primary 4-digit industry. Thus, treating industry as a fixed effect may be a mispecification of the model.

The balanced LRD panel is generally over-represented by large establishments. Thus, it is not simply small plants with little total output that are involved in industry classification switches. Abbott and Andrews (1988), report that primary 4-digit industry switches among plants in contiguous censuses account for over 3 percent of total output in both the 1972-77 and 1977-82 periods. For some 2-digit classifications, the average 4-digit industry had 10 percent of its output involved in switches. In short, the output effects of these switches on industry totals are significant in many industries.

3.1.3 Product Class Affiliations

These observed switches, in addition, are not simply the result of measurement errors associated with multiple output plants being reclassified from one industry to another because the "primary" output of the plant changes. Much of the "switching" activity involves adding or dropping whole product areas. Based on a comparison of matched establishments observed in both the 1981 ASM and the 1982 census, we found, based on some data developed as part of a study by the Department of Commerce's Bureau of Economic Analysis (BEA), that the gross outputs involved in switches averaged over 10 percent of total output in 1981 for both switches into and out of a product class.[11]

Tables 3 and 4 provide data on the percentage of product class output produced by plants that had production in the product class in 1981, did not produce in the product class in 1982, but did produce in one or more different product classes in 1982. This figure is termed the percentage of output that "switched out" of the product class. "Switched in" output is analogously defined also using the 1981 product class output as a base.

The Tables show that the distribution of the gross changes are dispersed and quite large. The average product class had over 10 percent of its output switched. In roughly 75 percent of the product classes, gross output attributable to switches in or out is more than 5 percent. In 5 percent of the product classes, over 70 percent of the output represented switches.

The net output effects of switches are substantially smaller, averaging less than 3 percent of total 1981 product class output. But as shown in Table 5, the distribution of the net changes shows relatively high values in certain product classes. The Table indicates that in over 10 percent of the roughly 1,350 usable product classes, the net effect of switches by matched establishments is greater than 5 percent.[12]

The data in Tables 3-5 reflect all 5-digit product groups available for analysis. One might object that this procedure overstates the problems by including a variety of miscellaneous product classes. This is indeed true for the net changes shown in Table 5. When we excluded all product classes ending in zero or 9, the miscellaneous categories, over 90 percent of the product classes showed the percentage of 1981 total output represented by net switches in the range between -3.5 and +3.5 percent. Nonetheless, for this set of product classes the percentage of total output subject to switches, although smaller than found in Tables 3 and 4, still averaged around 9 percent with a standard deviation of about 15. The phenomenon of large proportions gross output being associated with switches is not simply the result of poorly defined product classes.

These findings suggest that industry effects cannot be simply thought of as time-invariant effects. Moreover, switches are not simply the result of random (or nonrandom) measurement errors arising from problems with the SIC classification system. Research has not yet established if the probability of switches is greater at the time of ownership change than at other points in an establishments' history. Some evidence (McGuckin and Andrews (1988)) points this way. But managements do make economic decisions to reallocate a plant's productive capacity to new activities. This is true at the time of ownership changes. It is also true in day to day decisions as multiple product plants shift production in response to, among other things, changes in product demand. This suggests that for some problems at least, switches need to be treated as endogenous or an explained phenomenon.[13]

3.2 Implications for Aggregate Time Series Data

It is important to recognize that the work cited above has implications for aggregate analysis. The first implication is directly related to the discussion of the applicability of models of the type represented by equation (1). Aggregate analysis makes use of assumptions concerning the nature and homogeneity of individual economic agent's behavior. The evidence from LRD based research suggests that typical assumptions about the "representativeness" of aggregate observations may be inappropriate.

Aside from the modeling aspect, a second implication concerns the character of the observed time series.

Because of various processing considerations, most changes in industry output (or employment) associated with switches occurs in census years. One of the primary reasons for this is that as part of each census, a complete canvass of establishments is undertaken and an extensive company organization survey is conducted. Firms are asked to give a description of all products produced in their plants. The Census Bureau uses the new information to reclassify plants, so that plants are sent correct survey forms for the census. It is thus in census years that many of the switches are identified.[14]

The large portions of industry output that are subject to switches and the realities of processing imply that published aggregate output, employment, or other establishment based variables will contain discrete jumps between ASM and census years. Observation of these jumps led to the BEA project that collected the data underlying Tables 3 to 5. Recognition of the source of these jumps should provide information to improve the quality of available time series. It also should aid in the development of reconstructed time series that can be compared to unadjusted time series data obtained from the traditional cross-section aggregations of surveys and censuses. Such studies should yield important information for the interpretation of time series models.

3.3 Gross Flows

The LRD enables one to develop information on gross as well as net flows of economic variables such as job creation and entry. The opportunities for examination of measures of gross change has motivated a number of studies at CES. These studies collectively suggest that reliance on aggregate cross-sectional measures of net change may obscure important economic phenomena.

3.3.1 Job Reallocations

Recent work by Davis and Haltiwanger (1989) suggests that gross measures of job creation are important in the study of business cycles and other macroeconomic issues. Davis and Haltiwanger find that manufacturing employment contracted at a rate substantially less than the rate of gross job reallocations (the sum of job creation and destruction rates) in the 1972-86 period. The size of the gross reallocation rates relative to the observed net charges, (roughly 10 percent points greater) implies the existence of large worker flows across establishments that are masked by examination of net changes. Further, Davis and Haltiwanger find that gross job reallocation exhibits significant countercyclic time variation in contrast to the procyclical behavior of net job reallocations. The important point that Davis and Haltiwanger make is that gross measures of job creation and job destruction are important in the study of business cycles and other macroeconomic issues. As illustrated next, gross flow measures are also important in examining microeconomic issues.

3.3.2 Entry and Exit

The importance of the structure of a market in determining performance has long been emphasized. Until fairly recently, studies often relied on measures of market structure such as concentration ratios as an indicator of the likelihood of monopoly power. A concentration ratio measures the share of output produced by, for example, the largest four firms in a market.[15] In its simplest form, the theory suggests that the concentration ratio provides a measure of the ease of coordinating pricing policies by the largest firms in an industry.

There are many problems in using a concentration ratio alone as a measure of monopoly power. Among the most important is the long recognized importance of entry (or potential entry) as the ultimate constraint on firms that price above competitive levels. Until recently, little information has been available to construct measures of entry beyond simple net changes in numbers of firms.

One possibility for creating a dynamic measure of market structure based on gross entry and exit is to measure the number of large firms in a market who survive from one point in time to another. The theoretical justification for this measure is that it captures information about the turnover of competitors in a market. The measure is not new and the empirical tests reported here are only suggestive. But, they illustrate the possibilities and importance of longitudinal considerations in examining market structure.

The survival measure is developed for the roughly 450 4-digit industries in manufacturing for the years 1972-77, 1977-82, and 1972-82. The actual calculations included the 20 largest firms in terms of value of shipments in each census year in each industry. Thus, the number of survivors is simply the converse of the gross turnover of firms over the period. That is, we measure gross turnover simply as the total number of firms (20) less those firms that remain in the top 20. By construction, net turnover is zero.

Using the top 20 producers reduces the possibilities of misclassifications of small firms. For most industries the top 20 producers account for the bulk of industry output. They account for over 60 percent of industry output in 280 of the 450 available industries. Only in 55 industries did the top producers account for less than 40 percent of output. In fact, the average industry had roughly 75 percent of its output accounted for by the largest 20 firms in the 3 census years under consideration, 1972, 1977, and 1982.

The results of the calculations showed significant turnover among the largest firms. Table 6 shows that in a time span of as little as 5 years, the average industry replaced 8-9 top 20 firms. This figure implies a gross turnover rate of approximately 40 percent (8/20) for both the 1972-77 and 1977-82 periods. Measured across the 10 year interval 1972-82, gross turnover averages almost 60 percent with 11-12 of the top 20 firms replaced in the average industry.

These turnover rates do not suggest the widespread exercise of monopoly power. But, even with this large turnover, the market shares of the largest firms may be quite stable. This is the implication from the survival rate breakdowns by concentration class in Table 6.

Turnover, as expected, is greater when measured among the 20 largest firms than when measured among the 8 or 4 largest. Thus, reading across the rows in Table 6 one always finds the percentage survival rate increases with the number of firms in the initial size distribution, top 20, top 8, or top 4. However, there is little difference in turnover rates across concentration classes. Although the percentage of survivors was always smallest for the less than .4 concentration class, there is little difference between the two largest classes and the difference between these classes and the smallest also is not large. Moreover, regardless of the initial levels of concentration in the industry, the average industry lost approximately one firm from among the top four firms at the beginning of each period.

Although direct comparisons are not possible because of differences in procedure, Dunne, Roberts, and Samuelson (1988) also develop gross entry and exit rates for 4-digit industries.[16] A major difference in procedure relates to the treatment of ownership changes. Dunne, et. al. do not treat firms with ownership changes as new entrants unless the change alters the basic establishment structure of the firm in the market. They only consider as entrants firms bringing new capacity to the market. If new management takes over existing plants, this is treated as a name change.

In contrast, in this paper all ownership changes are treated as entrants and exits in calculating survival rates. If a new competitor is defined in terms of the capacity it brings to the market, then excluding "name changes" resulting from a merger or other ownership change makes sense. However, if, as suggested by the work on ownership changes cited earlier, new ownership brings new management and increased performance, then the "name" changes should count as entrants.

How much they should count is another question. One that cannot be decided on a priori grounds. Only with further empirical work relating performance and behavioral measures to survival rates and other measures of dynamic concentration will it be possible to sort out the proper measures.

The one sure conclusion that we draw from these studies is that panel data are necessary to make progress on these issues. While we have focused on the cross-section or across industry variation in turnover in this example, as with the work of Davis and Haltiwanger (1989) cited earlier, time series variations, reflecting shifts in demand, technological opportunities, or shifts in input prices, are likely to be important components of net and gross turnover. This is clearly the implication of the existence of merger cycles that have been identified with, among other things, industry shocks (see Blair (1989)).

4. CONCLUDING REMARKS

As suggested by the quote at the beginning of the paper, the available panels in the LRD permit a wide range of longitudinal studies. Here we emphasize two generic classes of studies that can be accomplished with panel data. The first is the so-called event study. In the examples cited, we show the importance of incorporating time varying effects in explaining establishment and firm behavior (both existing and new). Various studies at CES have shown distinct differences in firm performance following ownership changes. This work suggests that ownership changes need to be incorporated in models explaining establishment and firm behavior. Moreover, since the volume of mergers and other forms of ownership change varies greatly over time, these kinds of changes can have significant effects on aggregate time series data.

In this regard we also report on the large volume of establishment industry switches. These switches can generate jumps in aggregate industry output time series since, for a variety of reasons, the effects of such switches are largely accounted for in census years. In addition, since there is some evidence that these switches arise from ownership changes and other corporate events, their effect on aggregative output measures is not simply a processing or sample design consideration. Rather, it is a phenomenon to be modeled. At the very least, given the increased, number of mergers and acquisitions observed in the 1980s, an assessment of the effects of switches on aggregate statistics needs to be undertaken. We noted that current work at CES finds that gross job turnover measures have important implications for analysis of labor markets and business cycles. In addition, the importance of measuring gross flows was illustrated with a simple "dynamic" measure of market structure which exploited the LRD to obtain a measure of gross entry and exit.[17] In this respect, as well as in the event studies, we "get what we need."

REFERENCES

Andrews, Stephen H. and Thomas A. Abbott III (1988). "An Examination of the Standard Industrial Classification of Manufacturing Activity Using the Longitudinal Research Data Base, "*Bureau of the Census Fourth Annual Research Conference Proceedings.*

Blair, Margaret (1989). "Free Cash Flow and the Rise in Contests for Corporate Control," Brookings Discussion Paper in Economics.

Baldwin, John R. and Gorecki, Paul K. (1987). "The Dynamics of Firm Turnover," Economic Council of Canada Working Paper.

Baldwin, John R. and Gorecki, Paul K. (1989a). "Measures of Market Dynamics: Concentration and Mobility Statistics for the Canadian Manufacturing Sector," Forthcoming *Annales D' Economic et Statistique.*

Baldwin, John R. and Gorecki, Paul K. (1989b). "Job Turnover in Canada's Manufacturing Sectors," Statistics Canada Analytical Studies Branch Research Paper Series, No. 22.

Baldwin, John R. and Gorecki, Paul K. (1989c). "Firm Entry and Exit in the Canadian Manufacturing Sector," Statistics Canada Analytical Studies Branch Research Paper Series, No. 23.

Baldwin, John R. and Gorecki, Paul K. (1989d). "Mergers Placed in the Context of Firm Turnover," Department of Economics, Queen's University, and Statistics Canada, Draft.

Baldwin, John R. and Gorecki, Paul K. (1989e). "Dimensions of Labor Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover," Business and Labor Market Analysis Group Statistics Canada, Draft.

Caves, R.E. and Porter, M.E. (1978). "Market Structure, Oligopoly, and Stability of Market Shares," *Journal of Industrial Economics*, Volume XXVI, pp 289-313.

Caves, Richard E. and Porter, Michael E. (1980). "The Dynamics of Changing Seller Concentration," *Journal of Industrial Economics*, Volume XXIX, pp 1-15.

Davis, Steve J. and Haltiwanger, John (1989). "Gross Job Creation, Gross Job Destruction and Employment Reallocation," Forthcoming in Center for Economic Studies Discussion Paper.

Dunne, Timothy and Roberts, Mark J. (1986). "Measuring Firm Entry and Exit With Census of Manufacturers Data," Department of Economics, The Pennsylvania State University.

Dunne, Timothy and Roberts, Mark J. (1987). "The Duration of Employment Opportunities In U.S. Manufacturing," Department of Economics, The Pennsylvania State University.

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1988a). "The Growth and Failure of U.S. Manufacturing Plants," Department of Economics, The Pennsylvania State University Working Paper.

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1988b). "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries," *The Rand Journal of Economics*, Vol. 19, No. 4 (WINTER).

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1989). "Firm Entry and Post-Entry Performance in the U.S. Chemical Industries," Center for Economic Studies Discussion Paper, CES 89-6.

Gort, Michael (1973). "Analysis of Stability and Change in Market Shares," *Journal of Political Economy*, Volume 71, pp 51-63.

Gossack, Irvin M. (1965). "Towards an Integration of Static and Dynamic Measures of Industry Concentration," *Review Economics and Statistics*, Volume 47, pp 301-308.

Lichtenberg, Frank and Siegel, Donald (1987). "Productivity and Changes in Ownership of Manufacturing Plants," *Brookings Papers on Economic Activity*, pp 643-673.

Lichtenberg, Frank R. and Siegel, Donald (1989a). "The Effect of Takeovers on the Employment and Wages of Central-Office and Other Personnel," Center for Economic Studies Discussion Paper, CES 89-3.

Lichtenberg, Frank R. and Siegel, Donald (1989b). "The Effects of Leveraged Buyouts on Productivity and Related Aspects of Firm Behavior," Center for Economic Studies Discussion Paper, CES 89-5.

McGuckin, Robert H. (1972). "Entry, Concentration Change, and Stability of Market Shares," *Southern Economic Journal*, Vol. XXXVIII, No. 3.

McGuckin, Robert H. and Nguyen, Sang V. (1988). "Public Use Microdata: Disclosure and Usefulness," Center for Economic Studies Discussion Paper, CES 88-3.

McGuckin, Robert H. and Andrews, Stephen H. (1988). "The Performance of Lines of Business Purchased in Conglomerate Acquisitions," paper presented at the American Economic Association Meeting in Chicago.

McGuckin, Robert H. and Pascoe, George A. Jr. (1988). "The Longitudinal Research Data Base (LRD): Status and Research Possibilities," *Survey of Current Business*, Vol. 68, pp 30-37.

McGuckin, Robert H., Warren-Boulton, Frederick R., and Waldstein, Peter (1988). "Analysis of Mergers Using Stock Market Returns," Economic Analysis Group Discussion Paper, EAG 88-1.

Nguyen, Sang V. and Reznek, Arnold P. (1989), "Production Technologies, Economies of Scale, and Factor Substitution in Large and Small U.S. Manufacturing Establishments: A Pilot Study," Forthcoming in Center for Economic Studies Discussion Paper.

Table 1: Number of Establishments in the LRD for each year:

Year	number of establishments	Number of administrative record cases
1963	305,747	*
1967	305,611	118,622
1972	312,398	122,158
1973	73,460	-
1974	68,262	-
1975	71,145	-
1976	70,346	-
1977	350,648	144,648
1978	73,853	-
1979	57,559	-
1980	55,953	-
1981	55,045	-
1982	348,384	128,307
1983	51,619	-
1984	56,551	-
1985	55,128	-
1986	54,858	-

* There were no administrative record cases in 1963.

- There are no administrative record cases in the ASM.

Table 2: Variable in the LRD

Symbol	Variable	Availability*
ppn	permanent plant number	
id	identification number	
ind	tabulated industry code	
ppc	primary product class	
pisr	primary industry specialization ratio	
ppsr	primary product specialization ratio	
il3	status of establishment	
ei	employer identification number	
dind	derived industry code	
et	establishment type (0=ASM)	C
ar	administrative record (1=AR)	C
cc	coverage code	
sc	source code	
lfo	legal form of organization	C
st	state code	
smsa	smsa code	
cou	county code	
plac	place code	
va	value added	
vr	value of resales	
rcw	receipts for contract work	
msc	miscellaneous receipts	
te	total employment	
pw1	production workers:	March
pw2	production workers:	May
pw3	production workers:	August
pw4	production workers:	November
pw	production workers	(average)
ph1	personhours: January-March	
ph2	personhours: April-June	
ph3	personhours: July-September	
ph4	personhours: October-December	
ph	total personhours	

Table 2: Variable in the LRD (continued)

Symbol	Variable	Availability*
sw	total salaries and wages	
ww	wages: production workers	
ow	wages: other employees	
lc	total supplemental labor costs	
le	legally required supplemental labor costs	
vlc	voluntary supplemental labor costs	
cp	cost of materials, parts, etc.	
cr	cost of resales	
cf	cost of fuels	
ee	cost of purchased electricity	
pe	quantity purchased electricity	
cw	cost of contract work	
cpc	cost of purchased communications	A 77 & 82
fib	b.o.y. inventory: finished goods	
wib	work-in-progress	
mib	materials	
fie	e.o.y. inventory: finished goods	
wie	work-in-progress	
mie	materials	
tib	b.o.y. inventory: total	
tie	e.o.y. inventory: total	
nb	new building expenditures	
nm	new machinery expenditures	
ue	used capital expenditures	
bab	building assets - b.o.y.	A; after 73
mab	machinery assets - b.o.y.	A; after 73
bae	building assets - e.o.y.	A
mae	machinery assets - e.o.y.	A
br	building rents	A
mr	machinery rents	A
bd	building depreciation	A; after 76
md	machinery depreciation	A; after 76
brt	building retirements	A; after 76
mrt	machinery retirements	A; after 76
rbs	building repair	A; after 76
rm	machinery repair	A; after 76
m	material goods	C
mqpc	quantity produced and consumed	C
mqdc	quantity received and consumed	C
mc	delivered cost	C
pi	product code	C
pqp	product quantity produced	C
pqs	product quantity shipped	C
pv	product value shipped	C
pgit	quantity of interplant transfers	C
pvit	value of interplant transfers	C
pqpc	quantity produced and consumed	C
tvs	total value of shipments	C

* The variable is available for all years and all establishments except as noted

A = collected for ASM establishments only;

C = collected in census years only

b.o.y. = beginning of year

e.o.y. = end of year

Table 3: Percentage of Product Class Output in 1981 Switched Out of Product Class in 1982 Matched Plants*

```
                                                          Frequency of gross change
        Value of switched out output midpoint of              Cum                      Cum
                  percentage class                   Freq    Freq      Percent       Percent

   0 | ***************                               110      110        8.23          8.23
   1 | ****************                              119      229        8.90         17.13
   2 | ***************                               110      339        8.23         25.36
   3 | *************                                  99      438        7.40         32.76
   4 | **************                                102      540        7.63         40.39
   5 | ****************************                  212      752       15.86         56.25
  10 | *****************************                 220      972       16.45         72.70
  15 | *****************                             126    1,098        9.42         82.12
  20 | ***********                                    83    1,188        6.21         88.33
  25 | ******                                         47    1,228        3.52         91.85
  30 | *****                                          36    1,264        2.69         94.54
  35 | *                                              11    1,275        0.82         95.36
  40 | **                                             16    1,291        1.20         96.56
  45 | **                                             14    1,305        1.05         97.61
  50 |                                                 3    1,308        0.22         97.83
  55 | *                                               6    1,314        0.45         98.28
  60 | *                                               5    1,319        0.37         98.65
  65 | *                                               4    1,323        0.30         98.95
  70 | *                                               5    1,328        0.37         99.33
  75 |                                                 1    1,329        0.07         99.40
  80 |                                                 1    1,330        0.07         99.48
  85 |                                                 2    1,332        0.15         99.63
  90 |                                                 1    1,333        0.07         99.70
  95 |                                                 0    1,333        0.00         99.70
 100 | *                                               4    1,337        0.30        100.00
       --------+---------+---------+---------
              60       120       180
                   Frequency
```

* Gross switched out output is calculated by expressing 1981 output of plants producing in the product class that are producing in another product class in 1982 as a percentage of total 1981 product class output.

Table 4: Percentage of Product Class Output in 1981 Switched into Product Class in 1982 - Matched Plants*

```
                                                          Frequency of gross change
        Value of switched out output midpoint of              Cum                      Cum
                  percentage class                   Freq    Freq      Percent       Percent

   0 | *****************                             113      131        9.80          9.80
   1 | *******************                           145      276       10.85         20.64
   2 | ******************                            135      411       10.10         30.74
   3 | *****************                             127      538        9.50         40.24
   4 | **************                                 95      633        7.11         47.34
   5 | ******************************                226      859       16.90         64.25
  10 | ****************************                  209    1,068       15.63         79.88
  15 | ***********                                    85    1,153        6.36         86.24
  20 | *********                                      64    1,217        4.79         91.02
  25 | *****                                          37    1,254        2.77         93.79
  30 | **                                             17    1,271        1.27         95.06
  35 | **                                             17    1,288        1.27         96.34
  40 | *                                               6    1,294        0.45         96.78
  45 | *                                               8    1,302        0.60         97.38
  50 | *                                               7    1,309        0.52         97.91
  55 |                                                 2    1,311        0.15         98.06
  60 |                                                 3    1,314        0.22         98.28
  65 |                                                 2    1,316        0.15         98.43
  70 | *                                               2    1,318        0.15         98.58
  75 |                                                 1    1,319        0.07         98.65
  80 |                                                 0    1,319        0.00         98.65
  85 |                                                 2    1,321        0.15         98.80
  90 |                                                 0    1,321        0.00         98.80
  95 |                                                 3    1,324        0.22         99.03
 100 | **                                             13    1,337        0.97        100.00
       --------+---------+---------+---------
              60       120       180
                   Frequency
```

* Gross switched in output change is calculated by expressing the output of establishments producing in the product class in 1982 and in another product class in 1981 as a percentage of 1981 total output in the product class.

Table 5: Percentage of Product Class Output Due to Net Output Switches in Product Class in 1981 and 1982 - Matched Plants*

Absolute value of net output change due to switches midpoint of percentage class		Frequency of net change			
		Freq	Cum Freq	Percent	Cum Percent
0	**	792	792	59.24	59.24
1	******	109	901	8.15	67.39
2	****	80	981	5.98	73.37
3	***	58	1,039	4.34	77.71
4	**	49	1,088	3.66	81.38
5	****	84	1,172	6.28	87.66
10	***	65	1,237	4.86	92.52
15	*	29	1,266	2.17	94.69
20	*	13	1,279	0.97	95.66
25		8	1,287	0.60	96.26
30	*	15	1,302	1.12	97.38
35		5	1,307	0.37	97.76
40		4	1,311	0.30	98.06
45		4	1,315	0.30	98.35
50		4	1,319	0.30	98.65
55		0	1,319	0.00	98.65
60		1	1,320	0.07	98.73
65		2	1,322	0.15	98.88
70		1	1,323	0.07	98.95
75		0	1,323	0.00	98.95
80		2	1,325	0.15	99.10
85		1	1,326	0.07	99.18
90		0	1,326	0.00	99.18
95		0	1,326	0.00	99.18
100	*	11	1,337	0.82	100.00
	-----+-----+-----+-----+-----+-----+-----+-----+				
	100 200 300 400 500 600 700 800				
	Frequency				

* Net output change is calculated by expressing the difference between the output of establishments producing in the product class in 1981 and in another product class in 1982 (switches out) and the output of establishments producing in another product class in 1981 which produced in the product class in 1982 as a percentage of 1981 product class output.

Table 6: Surviving Firms Among Top Firms, Various Years Survivors in 1982 Among Top Firms in 1977

1977 Industry		Top 20		Top 8		Top 4	
Concentration Class		No.	%	No.	%	No.	%
Greater than	.6	11.7	58.5	6.0	75.0	3.3	82.5
Between	.4-.6	11.3	56.5	6.0	75.0	3.2	80.0
Less than	.4	9.7	48.5	5.6	70.0	3.0	75.0
Total		11.4	57.0	5.8	72.5	3.1	77.5

Survivors in 1977 Among Top Firms in 1972

1977 Industry		Top 20		Top 8		Top 4	
Concentration Class		No.	%	No.	%	No.	%
Greater than	.6	12.2	61.0	6.4	80.0	3.5	87.5
Between	.4-.6	11.7	58.5	6.5	81.3	3.5	87.5
Less than	.4	10.4	52.0	5.7	71.3	3.2	80.0
Total		11.8	59.0	6.2	77.5	3.3	82.5

Survivors in 1982 Among Top Firms in 1972

1977 Industry		Top 20		Top 8		Top 4	
Concentration Class		No.	%	No.	%	No.	%
Greater than	.6	8.8	44.0	5.1	63.8	2.9	72.5
Between	.4-.6	8.0	40.1	5.1	63.8	2.9	72.5
Less than	.4	6.6	33.0	4.1	51.3	2.5	62.5
Total		8.4	42.0	4.7	58.8	2.7	67.5

[1] The committee on Statistical Methodology argued that the essential feature of a longitudinal survey is that "from the beginning, there is a plan to elicit data from the future for each observational unit" (1986). The committee contrasted longitudinal surveys with surveys that support longitudinal analysis. The Longitudinal Research Database (LRD) was put, correctly, in the latter grouping. It was the only establishment panel among the 12 data sets studied.

[2] Most of the work with panel data has relied on data and models based on individuals.While many of the techniques applicable to individuals can be carried over to modelsof the behavior of firms and establishments, some new issues are involved. For example, the importance of ownership changes to establishment behavior has no obvious analogue for the individual. Analogies to the household as a unit of analysis are likely to be most apt. Nonetheless, the analogies are not perfect.

[3] Statisticians also emphasize the use of panel data in reducing collinearity and improving the precision of estimates in dynamic economic models involving lagged explanatory variables.

[4] The 1984 and new 1989 panels include, with certainty, the largest 500 firms in 1984 and 1989.

[5] While additions and subtractions to the sample are made each year to account for the formation of new establishments and the closing of existing establishments, there are various lags in the process, and some uncertainties are not resolved until census years.

[6] This number represents approximately 30 percent of the total ASM sample of establishments.

[7] Even here, however, the traditional emphasis on aggregate tabulations has an adverse effect on the available linkages.

[8] Current Industrial Reports data are not linked to the LRD. These reports contain yearly and sometimes monthly unit value data for many detailed SIC classifications. The CES has several specific projects working with these data and hopes to eventually be able to link CIR data more generally to the LRD.

[9] Some recent research suggests that prices differ across establishments and areas. Thus, the establishment may be a more appropriate level for deflation for certain research projects (see Abbott (1989)).

[10] Economic studies of this type are not widespread. One of the few areas where such work is common is in the finance literature. See McGuckin, Warren-Boulton, and Waldstein (1988) for an example of an "event" study using stock market data.

[11] For various reasons associated with the ASM sampling design, comparisons based on census and immediately preceding ASM years will overstate the annual switching rate. Even so, these numbers are large.

[12] These figures are calculated from data for the roughly 50,000 establishments sampled in the ASM panel. The output totals in 1981 are the product class totals published by BEA. The 1,337 product categories used in this study included all those with complete data and comparable definitions for each year. About 200 product classes were eliminated in the edit process.

[13] In the merger studies noted above, many industry and product class switches are associated with establishment ownership changes.

[14] Moreover, the ASM sampling design has resistance rules that limit establishment reclassifications in non-census years. Also, entrants are very difficult to track down and are often not observed directly until the organization survey is completed.

[15] The choice of the four-firm ratio rather than the three or two-firm ratio has its origins in the confidentiality protection rules employed by the Census Bureau.

[16] Their study includes all producers of a product, not just primary producers. They also exclude the smallest firms, those in aggregate accounting for less than one percent of total industry output.

[17] While we make no attempt to discuss these here, work in Canada on a database similar to the LRD is also suggesting that gross flow measures are extremely important for analysis of competition and export productivity, and labor job reallocations. Various work by Baldwin and Gorecki (1989b, 89c, 89d, 89e) suggests the Canadian Experience is very similar to the U.S. during the 1970s.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

MEASURING FIRM ENTRY AND EXIT WITH PANEL DATA

J.R. Baldwin and P.K. Gorecki[a]

ABSTRACT

The dynamics of the competitive process can be better understood with studies of entry and exit, growth and decline in the incumbent population, the effect of mergers, and the importance of the turnover process to productivity growth. Studies such as these require longitudinal data bases that measure firm performance over time. This paper describes the methodology used to build such a data base using material from the Census of Manufactures. The construction of a longitudinal panel from data that were not originally collected with this objective in mind is not easy. This paper outlines the difficulties and the choices that were made to resolve these difficulties. As more and more work is completed both in Canada and elsewhere on the dynamics of the competitive process, inter-country comparisons are increasingly made. This paper is meant to provide the reader of the accompanying studies on Canada with the means to evaluate these studies and to compare them, when appropriate, to the results for other countries that use other data sources.

KEY WORDS: Firm Entry; Firm Exit; Data Base Creation.

THE MAGNITUDE OF ENTRY AND EXIT

Introduction

The process of entry and exit of firms and plants has long been held to play an important role in the evolution and adaptation of industry to change. In the simplest of expositions, it is the act of entry and exit that serves to equate above or below normal profits to competitive rates. In other models, potential rather that actual entry serves to limit monopoly power. Once included under the rubric of limit-pricing models, this argument has been given theoretical elegance by contestability theory. The turnover process that results from exit and entry is also seen as a conduit through which new ideas and innovations are introduced.

Alternatively, entry can be portrayed as an interesting, but irrelevant, curiosity. One such view portrays entrants as fringe firms that swarm into and out of an industry without having much impact. References to the entry and exit process as "hit and run" leave the impression, intentional or otherwise, of an unstable fringe, which makes no contribution to such indicators of progress as productivity. Shepherd (1984), in a criticism of contestability theory, stresses that entry as an external force is usually a secondary factor to internal conditions within an industry in determining the strength of competition within an industry.

Despite the potential significance of the entry process, it is only recently that it has attracted much attention on the empirical side of the industrial organization literature. This newfound attention reflects a greater interest by industrial economists in the topic of market dynamics--how firms and industries behave over time and what effect this has on industry structure and behaviour.

Because of the dearth of empirical data on the entry process, the debate over the importance of entry remains unresolved. One of the difficulties of evaluating the importance of entry and exit has been a lack of longitudinal panel data that follow firms through time. The Canadian Census of Manufactures, as well as its counterparts in other countries, are designed to capture and report aggregate industry data at a point in time and until recently have not been able to follow the changes of individual micro units over time. Fortunately, the Canadian Census and related files contain individual establishment and firm identifiers that offered the potential of creating a longitudinal panel. In this paper, we briefly report some of the results of a project that used these data to measure the importance of entry and exit. More importantly, we describe how the data bases were created. The existence of identifiers does not by itself permit longitudinal studies--especially if the identifiers were not created with longitudinal studies in mind. All too often, such data bases are used for research studies without extensive documentation.

Short-run Rates of Change

In order to portray short-run change, rates of entry and exit are calculated annually from 1970 to 1982. The rates of expansion and contraction in continuing firms are also calculated for the same period.

Intra-industry dynamics are measured by focusing on changes in employment associated with each category. Employment changes are measured at the level of the consolidated firm. A greenfield entrant is defined as a firm that enters the manufacturing sector for the first time by building a plant. A closedown exit is a firm that leaves the manufacturing sector entirely by closing plant.

Rates are measured as the percentage of total sector employment in entrants and exits and as the ratio of the growth in expanding or decline in contracting firms as a percentage of total employment. Figure 1 compares the average annual expansion rates for greenfield entry to continuing firm expansion; closedown exits to contraction rates in declining firms.

On average, during the 1970s, entry accounted for 0.9 per cent of employment annually, expansion in the continuing sector for 7.8 per cent; exit accounted for 1.1 per cent and contraction for 6.3 per cent annually.

It is clear from this evidence that annual rates of entry are not large enough to suggest even moderate change is occasioned by entrants at birth.

[a] Business and Labour Market Analysis Group, Statistics Canada. J.R Baldwin is Professor of Economics, Queen's University, Kingston , Ontario, K7L 3N6; P.K. Gorecki, Senior Economist, Economic Council of Canada, K1P 5V6.

The Maturation Process for Entrants

The small values that are derived for the short-run or instantaneous entry and exit rates are not surprising. They confirm the casual impression that entrants rarely come to dominate an industry in their first year of operation. They might be used to support the view that entry is unimportant. That would be unwarranted at this stage. Such a determination must rely on more than the instantaneous rate of entry. Whether these new firms manage to grow in the longer period and displace existing firms, and how rapidly this occurs must also be examined.

Long-run measures of entry can be derived from the market share that has been accumulated by all entrants since an initial year. The total share of all entrants will increase over time because more cohorts are being added each year; but this tendency may be offset if the market share of existing cohorts declines. If, on average, each cohort adds n per cent to employment starting in period zero and then declines by a constant m percentage points per year, the maximum cumulative value that entry can have is in the n/m'th period.

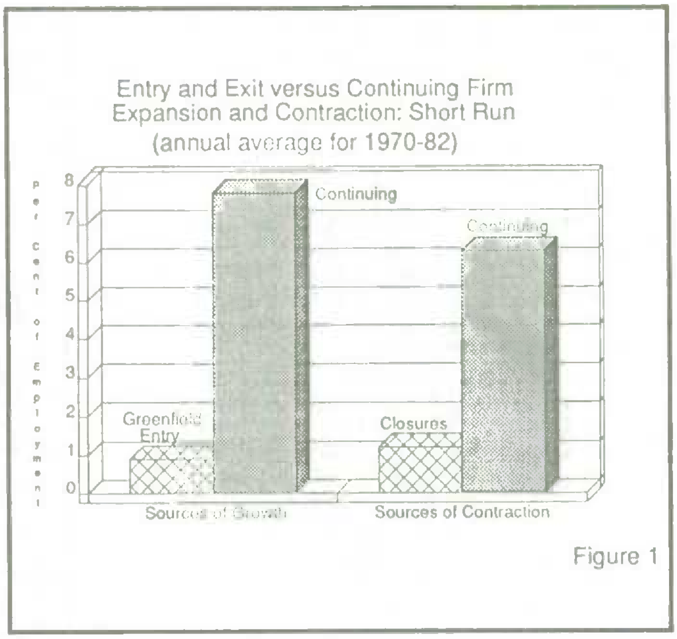

Entry and Exit versus Continuing Firm Expansion and Contraction: Short Run (annual average for 1970-82)

Figure 1

The long-run share of a particular cohort of entrants will depend on the exit rate, the average length of life and the growth rate subsequent to birth of all entrants in that cohort If entrants either experience a relatively short life due to high infant mortality rates or a relatively slow growth rate during adolescence then the long-run or cumulative impact of entry may be unimportant. On the other hand, surviving entrants may grow enough to outweigh the effect of exits and allow a cohort's share to increase over a substantial period of time. In this case, the cumulative effect of entry will be greater

In order to characterize the experience of surviving entrants in the 1970s, the data on entry to and exit from the manufacturing sector as a whole were used to calculate the share of each entry cohort as it matured. Data for each entry cohort from 1971 to 1980 were used and the average share, both in terms of number of firms and value-added, was calculated for each age class of each entry cohort. Because there is immediate exit from each entry cohort, the average percentage of all firms accounted for by each entry cohort declines continuously as the cohort ages. In contrast, the average value-added share increases throughout the period--some ten years studied here. The growth rate of surviving entrants then more than offsets the high death rate experienced by each cohort in the early years of its existence.

The average share value added share of a cohort along with the cumulative effects of succeeding cohorts are plotted in Figure 2. The average market share, using value-added, of each entry cohort from 1970-71 to 1980-81 is used for the starting point. The average share trajectory is then applied to each cohort. The resulting total market share captured by entrants is a representation of how the effect of entry accumulates on average. Over the decade studied, there is no downturn in an average cohort's share and, therefore, the cumulative effect of entry continuously increases. Despite their high mortality rate, entrants remain to make themselves felt as a group.

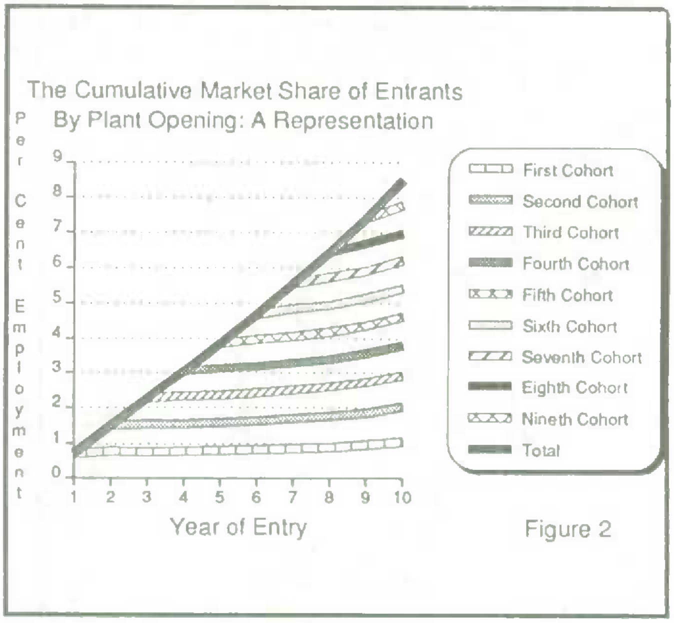

The Cumulative Market Share of Entrants By Plant Opening: A Representation

Figure 2

Cumulative Effects of Entry and Exit

Two six-year periods--1970-71 to 1975-76 and 1975-76 to 1980-81--and one eleven year period--1970-71 to 1980-81--are selected to examine longer-run entry and exit rates in the Canadian manufacturing sector. The long-run rates of change for each period are calculated by comparing the status of firms in the initial and terminal years. Thus, for the period 1970-71 to 1980-81, the entry rate is calculated as the 1981 employment in manufacturing firms that were not in the manufacturing sector in 1970 divided by 1970 employment in the manufacturing sector. This measure captures the cumulative effect of all entrants from 1971 to 1981 that were extant in 1981.

Figure 3 is a bar chart depicting the total cumulative contribution made to employment by firm entry and continuing firm expansion; continuing firm contraction and firm exit between the years 1970 and 1981 Entry added 10.9 per cent, and expansion 27.2 per cent to 1970 employment; exits led to a disappearance of 10.5 per cent of employment and contraction to 11.0 per cent of initial employment levels.

The results reported herein reflect only a portion of the research that has been conducted with the longitudinal Census data. Nevertheless, it serves to illustrate a conclusion that emerges from the work that has been done. While continuing or surviving firm activity is of substantial importance, the longer the time period selected, the more important the entry and exit process becomes relative to that of continuing firms. The importance of entry and exit only emerges in the longer run.

It is, therefore, important to be able to build longitudinal data bases that follow plants and firms over time. This is not an easy task. If the results of these exercises are to be evaluated, the manner in which the data were constructed needs to be fully developed. The remainder of this paper discusses the approach that has been taken to build the data bases used to measure entry and exit in particular and firm turnover in general (see Baldwin and Gorecki, 1989a to 1989g).

Entry and Exit versus Continuing Firm Expansion and Contraction: Long Run (cumulative 1970 to 1981)

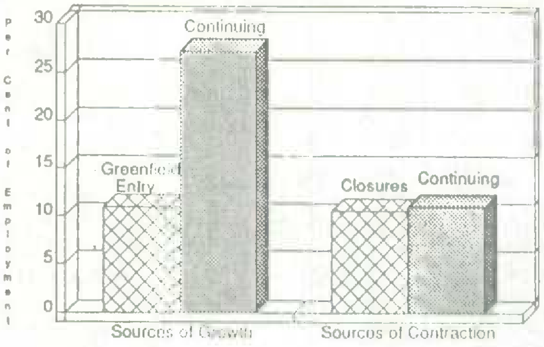


Figure 3

METHODOLOGY OF MEASURING ENTRY AND EXIT IN CANADIAN MANUFACTURING

Entry and exit can be defined as the emergence of new producing units and the disappearance of old units. Unfortunately, what is relatively easy to define in this manner is difficult to measure precisely. Varying interpretations can be placed on "new" and "old". Data may not be amenable to measuring the concepts desired. As a result, empirical work in this area needs to specify carefully the concepts being measured and the methods being used.

The previous section ignored all of these issues in presenting one set of results on the magnitude of entry and exit. The remainder of this paper describes the methodology used to generate manufacturing sector entry and exit statistics for the 1970s and early 1980s. It is part of a series of papers that discuss the dynamics of change in Canadian industry. It does not present detailed results; but selected excerpts are included for illustrative purposes.

In order to measure entry and exit, several questions must be answered. The first is the type of study for which the entry and exit measures are going to be used. Although the data bases discussed herein were created basically for studies of the competitive process, they were also used for job-change studies (Baldwin and Gorecki, 1989f, 1989g, 1990). Measures that are useful for one type of study are not necessarily valuable for another type. Second, there are broad conceptual issues that need to be resolved. What time period, what types of entry and exit, and what level of industry detail should be provided? Third, there are problems that arise during the actual measurement process?

The paper starts by discussing how research with different objectives may require different measures. It then examines the choice of time period, industry definition, and entry categories. Third, a broad overview of the data bases is given and a definition of the entry and exit categories actually used is provided. Finally, the detailed problems of implementation are discussed.

RELATING DEFINITIONS TO OBJECTIVES

Users of administrative and survey data have to proceed cautiously when they employ these sources for purposes that were not originally envisaged. This is especially the case when the appearance and disappearance of identification codes in these data bases are used to define births and deaths. Identification codes can appear and disappear for a number of reasons -- none of which may satisfy the particular definition of entry and exit that the researcher has in mind.

There are many different ways of defining a birth or death, since a firm is defined, not in a single dimension, but by a vector of characteristics. These characteristics include such variables as industry, ownership, country of control, size, and the location and number of plants. The multidimensional nature of the firm's characteristics would be unimportant if only one of those characteristics were required for defining births and deaths, or if all changed simultaneously. Neither is the case.

The nature of the research question determines the definition of a new or an exiting firm that is required. If research is directed at asking how the creation of new firms affects employment and job turnover in the first instance, then a greenfield definition of a new firm is the most appropriate; that is, the new firm should be one that has arisen because of the construction of new plant. This definition primarily depends upon the plant and employment status variables in the vector of characteristics that define a firm. A new firm is one that builds new plant, thereby generating new employment. A new firm that is simply a reincarnation of an old one under a new name should not be defined as a birth for studies that look at job turnover. Thus, mergers need to be excluded from the new firm definition used to measure job turnover or they should be treated as a separate category.[1]

Studies of the competitive process require a different definition of birth and death. If research is directed at evaluating the effect of new firm creation on competition, then births should be defined as the creation of new entities. In this case, births should include both entry via greenfield construction (a category that depends upon the plant status variable in the vector of firm characteristics) as well as entry by acquisition of existing plants (a category that depends upon the ownership status variable in the firm characteristics vector). The two different forms of birth should be distinguished, because they may have different effects on performance.

The wide range of interpretations that can be placed on the notion of entry and exit means that it is difficult to produce a single estimate that satisfies more than one purpose. Therefore, several data bases were constructed for the work reported in Baldwin and Gorecki (1989a to 1989g, and 1990). The following sections describe the nature of the conceptual issues that had to be resolved, the data bases used, and the categories selected.

CONCEPTUAL MATTERS

Decisions have to be taken with respect to such issues as the appropriate level of industry aggregation, the production unit to be used, the time period selected for measurement, and the categories of entry and exit to be employed. The appropriate choices in each of these areas are interrelated.

- Choice of Industry Level

Entry and exit can be measured either at the aggregate level of the manufacturing sector as a whole or at a finer level such as a 4-digit SIC industry. Since interest is usually directed in industrial economics at the extent to which entry and exit facilitates the equilibrating process, statistics at an individual industry level are required.

Nevertheless, entry and exit statistics at an aggregate level can be useful. First, it may be of interest to know how many outsiders to the manufacturing sector establish a new presence therein. Second, where the aggregate data are representative of individual industry level data, aggregate data usefully summarize the underlying trends at far less cost than is entailed in the creation of the individual industry series When plant entry is being measured, aggregate data will provide an adequate depiction of the amount of overall entry in individual industries. This is because a new plant in a particular 4-digit manufacturing industry is also an entrant to the manufacturing sector as a whole. Aggregate plant birth and death rates are, therefore, a potentially useful way of summarizing the underlying activity within individual industries.

The usefulness of aggregate firm entry and exit rates is more problematic. The number of firms entering the manufacturing sector need not be the same as the number of firms entering all individual 4-digit industries A firm may enter a particular 4-digit industry without being an entrant to manufacturing as a whole -- if it already existed in some other 4-digit industry. Aggregate firm entry rates will measure the amount of activity at the underlying industry level if most firm entry at the individual 4-digit industry level is done by firms that are new both to that industry and to the manufacturing sector as a whole. Whether this is the case is an empirical matter.

- Firm versus Establishment Data

Interest in the firm and plant turnover process for industrial economists centres on its role in affecting the evolution of industry profit, innovation, and productivity over time. Such considerations suggest that the appropriate unit of analysis is the firm rather than the individual production unit--the plant, factory, or establishment. It is the firm, not the plant, that makes the decision to enter or exit an industry.

On the other hand, plant entry rates are useful since they give a broad overview of the importance of all new plants that are created by both entering and continuing firms. It is this variable that may have the greatest influence on the equilibrating process that drives down supra-normal profits. For job-creation studies, it is the plant opening and closure process, rather than the new firm creation process -- which also includes a merger component -- that is relevant.

All this means that measures of entry and exit that capture both firm and establishment activity are useful in different contexts.

- Time Period

A choice has to made about the length of time over which entry and exit is to be measured. It can be estimated by comparing the status of firms and plants at two adjacent points in time using annual data, or by using endpoints that are further apart. In the latter case, the status of the firms in the interim is ignored. The first measures yearly rates of change; the latter investigates the cumulative effect of entry and exit over the particular time period. Taken together, the two measures can serve to evaluate the importance of entry as part of the competitive process, as well as the extent to which short-run job-turnover statistics capture mostly transitory or longer-run phenomenon.

- Entry and Exit Categories

The type of research also determines the nature of the entry and exit categories to be used. Job-turnover studies require classification systems that emphasize entry and exit of actual production units--plants or establishments. They require that firms and plants be distinguished. It is also important to divide firms into the continuing as opposed to the new and exiting segment in order to measure the relative contribution of each to job growth and decline. For this purpose, distinctions between new, exiting, and continuing firm and plant creation and destruction are required. Studies of the competitive process also require these breakdowns. In addition, such studies will benefit from the addition of a further category -- the entry and exit of firms via the acquisition and divestiture of plant.

- Resolution

The various issues that have been raised herein were resolved by adopting a set of measures that look at both the long and the short run. Both aggregate and disaggregate data on entry and exit are employed. The aggregate data are used to provide an overview of annual or short-run establishment entry and exit rates; they are also used to measure entry by new firms that created new plants. At the aggregate level, this series provides a reasonable indication of the amount of total activity at the underlying industry level. Disaggregated data are used for longer-run estimates. The most detail is provided for long-run estimates that capture the cumulative value of change. Establishments and enterprises are linked together to allow both new firm and continuing firm plant creation and merger activity to be measured.

THE DATA BASES

The necessity of measuring entry over different periods (the long versus the short run), at different levels of industry aggregation (individual sectors as opposed to the manufacturing sector as a whole), and for different producer units (firms versus plants) resulted in the creation of three different longitudinal data bases. Together, the three data bases allow entry and exit to be measured both at the individual industry level and for the manufacturing sector as a whole, between adjacent years and over longer periods, using firms and establishments--both separately and together.

The Longer-Run 4-Digit Industry Data Base

The first data base measures longer-run entry and exit by comparing the status of production units in 1970 and 1979. It provides detail on both establishment and firm status and links the two. Therefore, it can be used to measure both plant and firm entry and exit. It also allows continuing firm plant turnover activity to be measured so as to provide a standard of comparison for the entering and exiting firm sector. Finally, it measures activity at a detailed 4-digit Standard Industrial Classification (SIC) industry level.

Table 1

Plant and Firm Classification Matrix Used to Study Entry and Exit in Canada's Manufacturing Sector

	Firm Status		
Plant Status	Continuing	New	Dead
Divested	11	n.a.	31
Acquired	12	22	n.a.
Births	13	23	n.a.
Deaths	14	n.a.	34
Continuing	15	n.a.	n.a.
Transfer In	16	26	n.a.
Transfer Out	17	n.a	37

Definitions	Cell	
Entrants	22	Firms that entered the industry by acquiring one or more plants between t and t + n
	23	Firms that entered the industry by opening one or more plants between t and t + n
	36	Firms that entered the industry by transfering one or more plants from one industry to the given industry between t and t + n
Exits	31	Firms that left the industry by divesting one or more plants between t and t + n
	34	Firms that left the industry by scraping one or more plants between t and t + n
	37	Firms that exited the industry by transfering one or more plants out of the given industry to another between t and t + n
Continuing	11	Continuing firms that divested themselves of one or more plants between t and t + n
	12	Continuing firms that acquired one or more plants between t and t + n
	13	Continuing firms that built one or more plants between t and t + n
	14	Continuing firms that scraped one or more plants between t and t + n
	15	Continuing firms that owned at least one plant that existed in both t and t + n
	16	Continuing firms that transferred plants into of the given industry
	17	Continuing firms that transferred plans out of the given industry

n.a. = not appropriate

Plant births were defined as the appearance in 1979 of a plant in a four digit industry that had not been in that industry in 1970. A plant closure was equated with the 1979 disappearance from a four digit industry of a plant identifier that had existed in 1970. A firm entrant was defined as the appearance of a firm code in 1979 in a four digit industry that had not previously been attached to any other plant in the industry in 1970. A firm death occurs when the firm identifier attached to a plant in 1970 was no longer attached to any plant in the particular four digit industry as of 1979. Firms were defined as all commonly controlled establishments within a 4-digit SIC industry. It should be noted that plants or firms that enter after 1970 and die prior to 1979 are not captured in the entry or exit measures derived from the long term data base.

Because of the link between the plant and enterprise in the long-run data base, it is possible to measure a number of different entry categories. These are summarized in the plant and enterprise status matrix presented in Table 1. Cell identification codes, which are used subsequently to index variables, are also included in the table. The importance of the various categories can be measured using number of establishments, firms, shipments, employment or any other variable available from the Census of Manufactures.

The Annual Establishment and Enterprise Manufacturing Data Bases

The second and third data bases separately track the history of firms and establishments longitudinally on a year-to-year basis from 1970 to 1982. They are used primarily for short-term comparisons. Both these data bases define entry and exit at the manufacturing level as a whole. Births for establishments are defined as plants new to manufacturing. New enterprises are defined in a similar fashion. Plants that switch 4-digit manufacturing industries, or firms that do the same, are not defined as entrants in these data bases. Switches from outside the manufacturing sector--from wholesaling, for instance--are included as entrants in these data bases.

These data bases use an entry and exit classification scheme that is somewhat less comprehensive than was used for the long-run analysis. For the establishment data base, there was a three-fold classification establishments were either newly created, closed, or continued from year to year. In this data base, no account was taken of the owning enterprise; therefore, whether the plant was acquired or divested was not considered. The enterprise data base used a more detailed classification scheme. On the entry side, new firms were divided into those that did so by greenfield plant construction as opposed to acquisition, a similar distinction was drawn on the exit side.

The annual establishment and the enterprise data bases can be used to answer different questions. With the establishment data base, entry is defined as the creation of a new establishment, death as the closure of an existing establishment. Since the establishment data base covers all plant openings and closings for both continuing, new, and exiting firms, it is useful for measuring the extent to which plant turnover causes changes in employment. The enterprise base allows a distinction to be made between firm entry due to openings as opposed to entry due to acquisitions. It also allows the activity of new as opposed to continuing firms to be compared. Similar distinctions on the exit side are also possible using the enterprise data base. This data base can be used to answer questions about the dynamics of the competitive process at the firm level.

In order to comprehend more fully the meaning of the entry and exit measures provided by the three data bases, it is necessary to examine the definitions of establishments and enterprises that have been used and to describe more fully the categories that are employed. This is done in the next two sections.

ESTABLISHMENT AND ENTERPRISE DEFINITIONS

The measurement of entry and exit uses two basic units of production. These are the establishment or plant, on the one hand, and the enterprise or firm, on the other. The terms establishment or plant, and enterprise or firm are used interchangeably herein. Each of the terms needs to be carefully defined if the Canadian data are to be compared not only to those from other countries, but also to other data sets for Canada.

An establishment is defined by Statistics Canada as "usually equivalent to a factory, plant, or mill" [2] The focus of this paper is confined to establishments that are classified to the manufacturing sector.[3] The establishment is the basic statistical unit from which information is collected for the Annual Census of Manufactures.[4]

An enterprise is defined as all establishments in the manufacturing sector under common control.[5] An enterprise is thus a concept that does not necessarily coincide with the legal entity or what is sometimes referred to as the business or corporate entity. The relationship between the legal entity, the establishment, and the enterprise, is summarized by Statistics Canada

> "There is in fact an intermediate level of organization between the establishment and the enterprise, the legal entity. This is the ownership unit. Legal entities may be incorporated or unincorporated businesses, or individuals. One legal entity may own another legal entity; therefore, it is possible for an enterprise to control more than one legal entity, just as a legal entity may own more than one operating unit--an establishment" (Statistics Canada, 1983, p.24).[6]

Since an enterprise is defined as the unit that groups all commonly owned establishments, sub units can be created that combine all commonly owned establishments in a particular industry grouping (the 2, 3 or 4-digit industry). Thus, firm entry can be measured at the individual industry level or for the manufacturing sector as a whole. The finest level of detail (the 4-digit level) is selected for the longer-run 1970-79 data base The annual entry and exit rates are measured using the manufacturing sector as a whole. When results from one set are compared to another, it must be remembered that the two estimates need not be the same for the reasons discussed previously.

DATING ENTRY AND EXIT

Each establishment is assigned a unique identification number, the record serial number or RSN.[7] This number remains with the establishment as long as it is included in the Census of Manufactures.[8] Each enterprise is also assigned a unique identifier--referred to here as the ENT code.[9] Unlike an establishment's RSN, the enterprise code can change when one enterprise purchases another.

A birth of a plant or enterprise is defined as the appearance of a new identifier code. An exit is defined to occur when the code disappears. If the code continues over the period being studied, the plant or firm is defined as continuing. The short-run data base uses adjacent years from 1970 to 1982 to compare the status of establishments. The longer-run data base compares their status in the first and last year of the the 1970s.

Exit and entry are defined first by the status--the continuation, the discontinuation, or the creation--of the identification code of a plant and secondly, by the level of activity. Entrants are counted in the first year that the identifier appears and the employment or value of manufacturing shipments is positive; exits are defined to occur in the first year prior or equal to the actual disappearance of the identifier when the employment or value of manufacturing shipments falls to zero.[10] The latter criterion serves to exclude from the exit count those production units that, for some reason, had already ceased to be active participants. Use of the identification number alone in these cases, without consideration of whether production occurs, may cause the date of actual exit (based on production) to be estimated with a lag, since administrative systems and censuses are sometimes slow to purge themselves of defunct producers.

VALIDATION OF IDENTIFIERS

Entry and exit are measured by examining changes in enterprise and establishment identifiers. This section examines the reasons why these identifiers appear and disappear.

- The Establishment Code

Plant entry and exit is defined to occur with the appearance and the disappearance of an establishment code (the RSN). Whether this definition produces meaningful estimates of births and deaths depends upon the practice of the statistical agency in assigning establishment codes. The closure of an establishment is usually grounds for the retirement of a code; but there may be situations where continuing plants are reassigned

RSN codes--where the old code is dropped and a new one assigned. If continuing establishments are reassigned codes, then deaths and births will be overestimated.

Difficulties in this area arise because establishments, like firms, possess several characteristics. Some can change during the lifetime of a plant and cause the administrative coding system to assign a new plant number even though the plant has not closed. For instance, if changes in one of those characteristics--ownership--triggers a reassignment of a code, then death and birth will not correspond to the opening and closing of a plant.

The meaning of an establishment birth and death then depends upon the type of events that cause the statistical agency to reassign RSN codes to plants that have not shut down. The rule used by Statistics Canada is to discard codes in the case of a continuing plant and to assign a new one only if location, ownership, and name of the establishment all change simultaneously.[11] This rule precludes counting as an establishment death the situation where there has just been a change in the ownership or in the name of the plant.

The validity of entry measures that are developed depends upon the diligence with which Statistics Canada followed this rule. Two tests were employed to examine this. First, all cases were examined where a plant identifier that existed in 1981 had disappeared by 1985. All such plants were compared on the basis of their recorded names, ownership, and location to see if new plants could be found for the year 1985 that had similar names, ownership and location as one of the plants that had "died". Of the plants in existence in 1981, 12,076 had disappeared by 1985. These plants employed 206,668 workers in 1981. Of these only 10 plants, with 209 employees could be found that might have been coding errors; that is, only 10 new plants could be found in 1985 that appeared to be the same as a 1981 plant that had lost its plant identifier number. This suggests that the overstatement of the death rate, when measured using the disappearance of a plant identifier, is insignificant.

Administrative problems might also serve to cause the opposite type of error. If plant identifiers are reassigned when they should not be, there will be an underestimate of the 1981 death rate and the 1985 entry rate. The frequency of the opposite type of error was also investigated. Plant identifiers in 1981 and 1985 were examined to see whether two identical plant identifiers could be found, where the recorded name, ownership, and locations all changed. These plants should have had their 1981 identifier codes changed and new ones assigned by 1985. Some 18 such plants with total employment of 1298 were found. Once again the number of plants in this category made up an insignificant percentage of total exits.

Errors of the first type would cause an upward bias in the plant death and birth rates; errors in the second category would cause a downward bias in the plant death and birth rates. The errors in each case were small and essentially offsetting. Moreover, the errors were probably overstated since the identification of potential coding problems relied on mechanical computer routines and the existence of actual coding errors was not pursued further because the maximum potential error rate was already quite low.

In conclusion, because of the nature of the criteria used for the reassignment of the RSN plant identifier and the care used by Statistics Canada in following this criteria, the emergence of new establishment codes and the disappearance of old ones in the Canadian Census of Manufactures can generally be ascribed to "real" births and deaths. This is not the case in some other data bases where a change in legal entity is often sufficient to cause a code to be dropped and a new one to be created.[12] In this study, ownership and name of the plant can change, but as long as the location does not, there will be no change in the identifier and no false indication of a plant birth and death.

- The Enterprise Code

Enterprise identifiers (ENT codes) were used to track groups of establishments under common control. The same ENT identifier was assigned to all plants in manufacturing, logging, and mining owned by the same enterprise. This is not a code that corresponds to the legal entity, but one that is meant to relate to the concept of an enterprise that was discussed previously. Legal entity (BRID) codes do exist; new values of BRID codes are created and old ones discarded with a change in legal entity--such as an incorporation, an amalgamation a reorganization of establishments, or a change in ownership. Since the identity of the legal entity changes much more frequently than does the enterprise that controls the legal entity, the use of a legal entity (BRID) code can generate "false" births and deaths. Births in our various studies are considered to be false, if they involve only minor changes that fall neither into the entry by building new plant nor the entry by acquisitions categories that were defined previously.

Changes in the ENT code in the data base, by way of contrast, basically reflect only major changes in enterprise organization. An appearance of an ENT code in an industry should signify an entry by plant birth or by acquisition--where acquisition is broadly defined to include control changes which may not necesarily result in the merger of the facilities of the acquired firm with those of the acquiror. The disappearance of an ENT code should, likewise, be a death. As was the case with establishments, ongoing operations of enterprises are not supposed to have their codes retired and new ones assigned without good cause. However, in contrast to the case of the establishment, rules for reallocating ENT codes of ongoing enterprises are not as precisely specified. One reason for this is that the events that would have to be included in any definition are more complex. The rule as to name, location, and ownership used for a plant identifier change would not suffice.

ENT codes are supposed to change only when a major event takes place in the life of the enterprise. The extent to which this occurs in the data base was carefully examined. Not all categories where an ENT code appeared or disappeared had to be checked. Since each establishment was assigned both an RSN and an ENT code, attention was focused only on those establishments involved in either a firm entry by acquisition or a firm exit by divestiture (categories 22 and 31 in Table 1). This serves to eliminate those cases where the death or the birth of an enterprise was accomplished by the closure or the opening of an establishment (categories 23 and 34). Because of the care used by Statistics Canada in discarding and assigning establishment codes, the latter set of events were likely to have been associated with genuine enterprise deaths and births accomplished by plant closure or plant opening.

In order to evaluate the types of changes that occurred when an ENT code disappeared or appeared during the acquisition and divestiture of a plant, all establishments so affected were assigned to one or more categories. This served two purposes. The first was to evaluate the importance of the changes in the data base that were being classified as acquisitions and divestitures. The second objective was to isolate the number of cases where the ENT code had changed for only minor reasons, such as a name change that was not accompanied by a major event. Defining what is minor is more difficult than defining what is major. Therefore, a procedure of backward elimination was employed. Those cases where a major reorganization occurred were eliminated and then the residual category was examined.

Three events were defined as sufficiently major to rule out minor organizational changes. The first event was a change in the country of control of the enterprise that owned the plant in 1970 as compared to 1979.[13] The second event was finding that either the acquiring firm or the divesting firm continued throughout the decade.[14] In the former case, this meant the acquiring firm possessed a plant in some 4-digit industry in 1970 other than the one in which the acquired plant was located in 1979. In the latter case, this meant the firm, which exited an industry by divesting itself of plant, could be found in some other industry in 1979. The third event was defined as the presence of a horizontal merger.[15] This occurred when the firm that entered by acquisition did so by acquiring plants from more than one enterprise. It is unlikely that any of the major events outlined above could have occurred without there having been a major organizational change.

Each plant that was acquired or divested was categorized on the basis of the major event categories. The categories are not mutually exclusive, so a plant could be placed in more than one category. The importance of a category for entrants is measured as the ratio of the sum of the 1979 shipments of all plants in all industries contained in that category, divided by 1979 shipments of all acquired plants of entering firms. Importance of the various categories using plants of firms that exited by divestiture is defined similarily but uses 1970 shipments to measure importance.

Almost half of 1979 shipments of acquired plants of entering firms were in plants acquired by firms that possessed plants in another Canadian industry in 1970. These were diversifying mergers and not a minor form of corporate reorganization. Some 43 per cent of shipments in plants acquired by new firms were affected by a country of control change. Horizontal mergers within the acquisition and divestiture category were less important. Around 13 per cent of shipments in 1979 were in plants that were merged with other plants within the same industry at some time during the decade, and were also part of the acquisition process that brought new firms into an industry. After all the major event criteria were applied, only 25.3 per cent of shipments were not involved in a take-over by an existing Canadian firm, a country of control change or a horizontal merger.

When a similar exercise is performed for exits, it was found that the percentage of cases where divestiture was not accompanied by a horizontal merger, a country of control change, or the continuation of the divesting firm in another industry was somewhat larger than for entrants--some 35.1 per cent.

The two sets of plants in each of the residual categories do not entirely overlap. When the major event criteria were imposed simultaneously on both acquired and divested plants, there were only 8.6 per cent of all plants with about 9.5 per cent of shipments remaining that might not have been involved in a significant reorganization.

The plants in the residual category were checked manually. Ultimately, 3.4 per cent of the original establishments with 1.6 per cent of employment turned out to have involved a minor change in enterprise status like a name change. Reclassification of the group from acquisition and divestiture (categories 31 and 22) to continuing (category 15) had no effect on the importance of these categories as is reported in Tables 3 and 4.

In conclusion, an examination of the different methods of entry by acquisition and exit by divestiture confirms that this category is not misspecified. Because of the way in which new firm identifiers are issued in many administrative data bases, there is always the possibility that the phenomenon being measured is not associated with a major change in control or operating group structure.[16] Corporate reorganizations that result in a new legal structure but no change in ownership or operating structure or policies can occur for a number of reasons--for example, tax reductions. The validation checks of the enterprise identifiers that were carried out indicate that changes therein capture important economic events. They are not mere name changes, minor corporate reorganizations, or coding errors.

IMPLEMENTATION PROBLEMS: GENERAL ISSUES

The broad conceptual issues as to time horizon, industry detail, and the entry and exit categories to be adopted are relatively straightforward to resolve. More difficult are the problems associated with the peculiarities of individual data bases that make precise measurement a problem. This section provides a broad overview of some of these problems and their severity for the data taken from the Census of Manufactures that were used herein. A more detailed description of the specific resolution of each problem is reserved to subsequent sections where the difficulties experienced with each data base are presented.

- Coverage

The value of entry and exit statistics produced by a data base will depend upon the comprehensiveness of the coverage provided by the data base. Data bases like the Dun and Bradstreet records used by Birch (1979, 1981) and the U.S. Small Business Administration (1984) are incomplete--being constructed only from the records of those companies that wish to be placed on these files for credit rating purposes. Other data bases, like the ones constructed by Storey[17] and his colleagues in the U.K. are built from different sources, none of which purports to be a complete census.

Use of the Canadian Census of Manufactures to measure entry and exit overcomes these problems in large part. The Canadian data, cover all firms in the manufacturing sector, and are collected by the official statistics agency. These data embody the professional expertise and extensive coverage associated with the collection of national censuses.

Problems can also arise for longitudinal data bases, not so much because coverage is incomplete, but because it is not current or because it changes over time. This is often the result of there being a lag in adding new firms to a data base or in purging it of firms that have exited. Sudden bursts of activity to catch existing firms that may have been overlooked or to purge the files of defunct producers can generate a spurious level of measured entry and exit for a particular year.

Because the Canadian census is annual, it is generally not affected by these problems. An effective method for finding new plants and firms exists--through the use of administrative tax files. Moreover, failure of a previously existing producer to file a census return is followed up by trained personnel to ascertain the status of the firm or plant. The Canadian census is, therefore, generally both comprehensive and current. Meaningful annual rates of entry and exit can be derived therefrom. There will be some lags and omissions but they will be minimal compared to alternate sources.

The Canadian census data are not completely immune from the problem of changing coverage over time. A change in coverage occurred in the mid 1970s. However, information exists that allows the precise effect of changing census coverage to be estimated.

The 1970-79 data base that has been developed to measure entry and exit in the long run should not be affected by this problem since many of the missed entrants in the mid 1970s will have been restored by 1979. However, the establishment and enterprise data bases that are used to measure annual rates of entry were affected and modifications were required to handle this problem. These are discussed in a subsequent section.

- Sample Choice

The advantage of using an official census is based on the extensive coverage such data provide. The disadvantage is that it can be extremely costly to employ all records for the analysis. Moreover, it must be remembered that not all records are of equal quality.

An establishment that is surveyed directly by Statistics Canada for the Annual Census of Manufactures may receive either a long-form or a short-form questionnaire. The distinction between the two is:

> "The long-form is a fully detailed questionnaire sent to establishments with shipments above minimum sizes which vary by province and by industry and from year to year, designed to capture all but a small percentage of the shipments of the industry. In 1975 long-forms accounted for all but 4.1 per cent of the value of shipments of goods of own manufacture of the manufacturing industries. The short-form is a simplified, abbreviated questionnaire, bearing a closer resemblance to a typical income statement It is sent to small manufactures whose shipments fall below a minimum size"(Statistics Canada, 1979, p.10).

Some very small plants do not receive either a short- or long-form. Data for these small plants are taken from taxation administrative records in place of mailed questionnaires. In the late 1970s and early 1980s, both types of small establishments[18] accounted for 5 per cent or less of all manufacturing shipments: 2.0 per cent in 1970, 4.1 per cent in 1975, and 3.4 per cent in 1982.[19] In contrast, such establishments accounted for 40 per cent of all manufacturing sector establishments in 1970, 50 per cent in 1975, and 53.9 per cent in 1982.[20]

Understanding the difference between large and small establishments is important because it is sometimes opportune, for cost reasons, to work with only a subset of all establishments when entry and exit is measured Moreover, the creation and disappearance of small establishments may be sensitive to the diligence used in finding these small establishments. This, in turn, can vary year by year depending upon the budget constraints faced by the statistical agency and official concern about the paper burden imposed on smaller firms.

In this study, typically only long-form establishments are used for the reasons described above and because the use of the long-form sample permits more characteristics of entrants to be measured consistently. This is because the long form data contain more detailed information on plants' activities and because certain concepts, such as value-added, are not defined in the same way for long and short-form establishments.[21]

The impact of using this sample was investigated by comparing entry and exit rates using the universe of census establishments and just the long-form sample. For this purpose, the longer-run data base, with an initial year of 1970 and a terminal year of 1979 was employed The long-form sample yields a much lower rate of entry and exit than the entire sample when numbers of plants and enterprises are employed; but its use does not greatly affect the estimate of these rates when measured in terms of employment or shipments.[22] This is discussed further below. Small establishments are numerous but account for an insignificant percentage of total employment.

The same reasons that led to the selection of only long-form establishments also determined the choice of enterprises that only owned long-form establishments for the short-term data bases. An enterprise is defined in the Census of Manufactures in terms of the establishments it owns. The establishments of larger enterprises typically are classified as long-form; small enterprises as short-form.

Adoption of the long-form sample does create certain addition problems. The cut-off between a short- and a long-form establishment was changed drastically in 1975. This does not create any major problems for the longer-run data base. It would if a comparison was being made of the periods 1970-75 and 1976-80, because there would be slightly fewer entrants in the latter period. It creates more of a problem for the measurement of annual rates of entry and exit in that a discontinuity develops in the middle of the period. Discussion of this will be found in a subsequent section.

- Units of Measurement

The importance of entry can be measured either in terms of numbers of establishments and enterprises, or their outputs and inputs. Both sets of measures are used. Entry and exit rates calculated using numbers reveal whether entry and exit is easy; when calculated using an output or input size measure, they indicate whether it is important. Both shipments and total employment (wage and salary earners) are used to measure size. Shipments is the most logical measure to use for studies of the competitive process because it indicates what share of the market entrants are able to capture. Employment is also used to provide information on the contribution of entry and exit to job turnover.

Throughout, employment is derived from the total activity statistics available from the Census.[23] It is reported by the Census as an annual equivalent. For example, if a plant employs 60 workers per month for six months, this is recorded as 30 person-years. In some cases, this procedure might produce a downward bias in the estimates of entry and exit--for 60 not 30 people are affected by the exit of the above-described plant. This, in turn, would affect calculated rates of entry and exit because, presumably, the employment of continuing plants, which forms the denominator of this calculation, will not be affected to the same degree by this factor. One approach would be to assume that entrants and exits are distributed uniformly across the year--that they have an average life of half a year. All raw employment figures for entry and exit would then be doubled.[24]

This is not the practice that has been followed herein. It is felt that there is enough of a reporting lag in the Census that employment totals for the first and last reporting year of an establishment are for essentially a full-year's operation. This was tested by examining employment in enterprises that exited, both in the year of exit and the preceding year. The differences were relatively minor and certainly not of an order of magnitude of 100 per cent, which the doubling rule would imply.

- Definitional Nuances

After the categories to be measured have been determined, problems of implementation remain, because there are some cases where alternative definitions can be used to measure a particular entry and exit category. Two questions were examined carefully. The first was whether plants that are switched from one industry to another should be counted as establishment deaths and births. The second was whether firm entry and exit categories overlap one another.

Plant reassignment as entry Establishment entry is defined as the appearance of a new plant. A new plant may appear in a particular 4-digit industry because it did not exist in the Census of Manufactures. It could also be that it existed previously in some other industry but was switched to the new one. An establishment is assigned to an industry on the basis of the commodities that it produces. As a plant's commodity output changes, the industry to which it is assigned by the Census may change--though this is usually done with a lag in order to ascertain whether the change in output of the plant is a permanent phenomenon. Switches occur because plants that were previously concentrating on products assigned to other industries are now more heavily concentrated on products in the industry in question

The appropriate treatment that should be accorded plants that were reassigned from one industry to another is difficult to specify. The reassignment of existing plants from industry M to industry N causes the Census to shift the entire employment from M to N but is not generally associated with the creation of new employment in N equal to the total work force of the reassigned plant. Therefore, entry measures that include this form of entry in N, at first glance, appear to overstate the job creation and destruction associated with entry and exit. This argument would suggest that, for job turnover studies, switches should be excluded.

Table 2

Percentage of Industries with Non-Zero Observations the Various Entry and Exit Categories Across 167 4-digit Canadian Manufacturing Industries: 1970-1979

		Firm Status					
Plant Status		Continuing		Entrant		Exit	
		All Obs	Long Form Sample	All Obs	Long Form Sample	All Obs	Long Form Sample
a)	Divested	32.9	32.3	-	-	91.0	91.0
b)	Acquired	52.7	52.7	88.6	88.6	-	-
c)	Birth	74.8	73.6	99.4	94.0	-	-
d)	Closed	74.9	74.8	-	-	97.6	96.4
e)	Continuing	100.0	100.0	-	-	-	-

Notes: 1) See Table 1 for definition of the plant and firm status. All entry and exit categories are measured for the period 1970-1979.

2) Plant switches are not considered when calculated category c or d.

Source: Special Tabulations: Business and Labour Market Analysis Group. Statistics Canda.

On the other hand, for studies of competition, switches are important because they bring new participants into the industry and thus they should be included.

The matter was resolved by measuring the importance of switches using the long-run data base. For the short-run data bases, the switches are probably less of a problem. The data bases that are used to measure annual rates of entry consider entry and exit only to the manufacturing sector as a whole. Plant switches from one 4-digit industry to another within manufacturing are not a problem here. However, at the aggregate level being used, entry and exit switches may occur if plants are reassigned from manufacturing to wholesaling. In the short-run data bases, switches are included as entrants and exits and no attempt was made to measure their precise magnitude.

Overlap in Firm Entry Categories .When entry and exit is defined in terms of plant numbers, there are few problems of overlap. Plants fall exclusively into one or other category.The overlap problem is potentially more serious when the number of firms is used to measure entry. Firms may enter by building new plant, by acquiring new plant, or by doing both. Continuing firms may build new plant, divest plant, and acquire plant. This creates several potential difficulties. The percentages in various categories no longer sum to 100. The comparisons of entry intensity across industries then can be influenced by differences in the intensity of multiple category activity. The importance of this problem was investigated using the longer-run data base.

IMPLEMENTATION ISSUES: SPECIFICS

The previous sections of the paper have described and discussed in a general fashion the definitions and choices made in generating three data bases for studying various aspects of entry and exit. The following sections examine each of the data bases in greater detail. Emphasis is given to the way in which the implementation problems were resolved.

The Longer-Run Data Base

In the short run, the cyclical and stochastic components of firm growth and decline tend to overwhelm the structural trends. This is also the case with entry and exit. Because it was felt that the importance of entry and exit would emerge only in the longer run, the long-run data base provides the most detail. Entry and exit are measured at the 4-digit industry level. All of the categories in Table 1 are used.

- Sample choice

As has been indicated, the extent of entry and exit can be estimated using the entire universe of firms and establishments or the reduced long-form set. There are advantages to using only the reduced set of long-form establishments. But before this sample is used extensively, it is important to evaluate the effects of doing so

Table 2 contains the percentage of all 4-digit industries for which there were non-zero observations in each of the entry and exit categories. The coverage ratios are presented both for the entire set of establishments in each industry and for only the long-form sample. It is evident that the choice of the long-form sample does not greatly affect coverage.

Table 3 contains two estimates of the importance of the various entry categories using both number of establishments and the value of shipments. Table 4 contains estimates of the importance of the exit categories using the two samples. The first estimate in each case uses the entire set of observations; the second uses the long-form sample. The importance of an entry or exit category is measured relative to the totals for the set used--all observations in the first case, only long-form observations in the second. The estimates presented in Tables 3 and 4 are the average of the importance of each category taken across 167 4-digit industries.

It is evident that the use of the long-form sample affects the importance of entry and exit when numbers of establishments are used; but it has much less of an effect when shipments are used. Thus, the long-form sample may be employed to measure the shipment values affected by entry and exit without great distortion This conclusion also applies to other measures of input or output.[25]

- Entry of completely new as opposed to reassigned plants

Since establishment entry and exit can be defined either inclusive or exclusive of plants that have been switched from one industry to another, the magnitude of the plant switching category was investigated. In order to do so, all continuing establishments that were assigned to an SIC code in 1979 that differed from that assigned in 1970 were defined as entrants in 1979 to and exits in 1970 from the relevant 4-digit SIC industry by plant switching. Plant switches were divided into two categories: those attached to entering

Table 3

Importance of Categories for Entry Between
1970 and 1979 in Canadian Manufacturing
Using Alternate Data Sets:
(calculated as the mean across 167 4-digit industries)

Category	Share of Number of Plants		Share of Shipments	
	Total Sample	Long Form Sample	Total Sample	Long Form Sample
All	100.0	100.0	100.0	100.0
All Entering Firms by				
1) Plant Birth(23)	36.9	18.8	14.4	11.5
2) Acquisition(22)	6.5	8.7	10.4	10.7
3) Plant Transfer(26)				
a) no change in plant ownership	3.5	4.7	3.3	3.5
b) change in plant ownership	0.6	0.9	1.0	1.1
All Continuing Firms				
4) Continuing Establishments(15)	46.8	59.2	63.0	65.0
5) Acquired Plant(12)	1.6	2.2	2.8	3.0
6) New Plant(13)	3.6	4.6	4.2	4.4
7) Plant Transfer(16)	0.5	0.7	0.9	0.9

Note 1) for definitions of categories, see Table I and the text.
2) the importance of the various entry categories is measured as the number or shipments of plants owned by firms in a particular category as a percentage of all plants or all shipments in an industry.
3) the mean is taken across all industries --including those that have a value of zero in a particular category.

Table 4

Importance of Exit Categories in
Canadian Manufacturing Industries: 1970-79
for Alternate Data Sets
(calculated as the mean across 167 4-digit industries)

Category	Share of Number of Plants		Share of Shipments	
	Total Sample	Long Form Sample	Total Sample	Long Form Sample
All	100.0	100.0	100.0	100.0
All Exiting Firms				
1) Plant Death(34)	32.4	24.6	14.1	13.3
2) Divestiture(31)	8.5	10.0	12.5	12.7
3) Transfer(37)				
a) no change in ownership	3.8	4.3	3.4	3.5
b) change in ownership	0.6	0.8	1.3	1.3
All Continuing Firms				
1) Continuing(15) Establishments	50.6	55.3	62.9	63.4
2) Divested Plant(11)	0.5	0.6	1.1	1.1
3) Closed Plant(14)	3.3	3.8	3.7	3.8
4) Transfer(17)	0.4	0.5	0.8	0.8

Note:1) for definition of categories, see Table 1 and the text.
2) the importance of a category is defined on the basis of the number of plants or the shipments of plants in that category as a percentage of all plants or shipments.
3) the mean is taken across all 167 industries.

firms and those attached to continuing firms. In the former case, the plant switch brought a new firm into an industry. In the latter case, the firm, whose plant was reassigned to a new SIC, already possessed a plant therein. A firm's status--new as opposed to continuing--depends on its possession of plant in a particular 4-digit industry.

Tables 3 and 4 also contain estimates of the importance of entry and exit via switching. The rate of new firm entry via switching was 4.6 per cent using shipments and the long-form sample. This rate is not greatly affected by the sample chosen. The rate at which new firms are brought into an industry by plant switches can be broken into two subcategories. The first (row 3b, Table 3) are those that involved a change in plant ownership (1.1 per cent of total industry shipments using the long-form sample) and which might be included in entry by acquisition (category 22). This group is about 10 per cent of the entry by acquisition category that does not include plant switches (row 2, Table 3). The second category (row 3a, Table 3) consists of those plants that did not involve a change in plant ownership (3.5 per cent of total industry shipments). These might be included in the entry by new firm new plant class (category 23). Their shipments are equal to some 30 per cent of the new firm entry by plant building category that does not include switches (row 1, Table 3).

The result for exits mirrors that on the entry side. Switches that do not involve a change in ownership can increase the firm exit rate by plant closure by about 30 per cent. Plant switches by continuing firms are also important relative to new plant creation by continuing firms. They account for 0.9 per cent of 1979 shipments (row 7, Table 3) compared to de novo plant share for continuing firms of 4.4 per cent on average (row 6, Table 3). In conclusion, switches cannot be ignored since they have the potential to substantially affect the calculated long-term entry and exit rate.

- Overlap in Entry and Exit Categories

In order to investigate the extent of this problem, the number of establishments and the number of firms in the various entry categories were estimated for a reduced 141 industry sample – a sample that was used for regression analysis of entry. (Baldwin and Gorecki, 1987) Only long-form establishments were used. The results are reported in Table 4.

Across 141 4-digit manufacturing industries, an average of 24.6 firms per industry had entered by 1979, 4.9 by acquisition, 21.7 by de novo plant building.[26] Therefore, of the 24.6 entrants, 2 on average entered over the period 1970-79 by both acquiring plant and building new plant. In terms of exits, on average 38.3 of the existing firms as of 1970 exited over the decade, 7.2 by divestiture and 33.2 by scrapping. Thus of the 38.3 exits, about 2.1 on average exited over the period 1970-79 by both divestiture and scrapping of plant.[27]

In the continuing firm category, there were 50.3 firms on average--that possessed plant in the industry in the initial and terminal years. There were 49.8 owning plants that stayed in the industry over the decade. 1.6 that divested plants and 3.7 that scrapped plants. The sum of the subcategories (55.1) is about 10 per cent greater than the number that continued (50.3). Roughly the same overlap exists on the entry side when the number of continuing firms in 1979 is examined.

Measuring Entry and Exit in the Short Run

As was indicated above, two data bases were created to measure short-run entry and exit to the manufacturing sector. The first tracks establishments annually through the period 1970-82. The second tracks enterprises year by year over the same period. The short-run data bases measure entry and exit only at a high level of industry aggregation--the manufacturing sector as a whole.

A number of problems arose when short-run entry and exit were measured. These are discussed in the following two sections. The first deals with the establishment data base; the second deals with the enterprise data base.

The Annual Establishment Data Base

- Choice of Sample

Entry and exit data can be generated using all establishments, just long-form establishments, or just short-form plants. It was decided to use only long-forms because, amongst other things, the constantly varying coverage of short-forms would give rise to specious entry and exit[28] -- especially in the case of the measurement of annual rates of entry and exit. In a previous section, it was demonstrated that long-form data closely proxy the results of the total census for the longer-run period from 1970 to 1979--at least when entry is measured by the amount of shipments or employment affected.

For the short-run data base, the use of long-forms alone as a sample criterion is inadequate. The line of demarcation between short- and long-forms changed over time. Because of this, the use of long-form data alone would produce some changes in entry and exit purely as a result of reclassification. This problem was resolved by taking as the longitudinal establishment sample all establishments that completed a long-form on at least one occasion. An establishment then is classified as entering in a particular year, because it made its first appearance in that year and either was already a long-form, or eventually became a long-form at a later date.

This technique serves to reduce but not to eliminate the problems that shifting boundaries between short- and long-form establishments produce. It essentially smooths out the fluctuations by eliminating the most volatile component -- establishments just at the boundary. Since the boundary changes are generally small, this is sufficient most of the time; but, there are two occasions when major changes in census coverage occurred. For these instances, corrections in the estimates of entry and exit were required.

- Major revision in long-form coverage in 1975

The cut-off between a short-form and a long-form experienced a major revision in 1975.[29] During the early 1970s, Statistics Canada raised the cut-off point slowly to maintain approximately the same percentage of establishments in each category. But in 1975, the cut-off point was increased dramatically in order to reduce respondent burden for smaller manufacturers. As a result, the percentage of short-form establishments increased from 36.1 per cent in 1974 to 50.1 per cent in 1975. There was no subsequent increase in the percentage of establishments in the short-form category of a similar magnitude, though the percentage of short-forms drifts slowly upward over time. By 1983, it was 54.9 per cent of all establishments, as compared to 50.1 per cent in 1975. Over the same period, the percentage of employees in short-form establishments increased slowly from 7.6 to 8.7 per cent.

The reclassification of the boundaries between long- and short-forms in 1975 will have less of an effect on estimates of entry and exit with the use of the modified long-form sample adopted here.[30] This is because establishments that entered in 1975 as short-forms, but that eventually grew to become long-forms--albeit a harder task after 1975 because of the higher cut-off point used to define long-forms--will still be caught. However, it does not completely eliminate the problem. Those establishments that would have made the transition from a short- to a long-form under the pre-1975 definition of a long-form, but do not do so under the new definition, will be missed.

That there is some reduction in measured entry because of the 1975 change is evidenced by the increase in the average size of entering establishments that occurred subsequently. Entering establishments averaged 20 employees per establishment between 1970-1 to 1972-3, but 28.1 employees per establishment between 1975-76 and 1980-81. The increase in plant average size occurred abruptly at the time of the reclassification of establishments between the long-form and short-form categories in 1975.

In order to calculate the effect of the 1975 redefinition on the estimated entry rates, the distribution of entrants in 1973-74 was truncated by removing the smallest entrants until the average size of those remaining was equal to the post-1975 size of the average entrant. On average, this required removing 32.1 per cent of entrants accounting for 4.5 per cent of employees of all entrants. This is the estimate of the percentage reduction in the pre-1975 entry figures required to make them comparable to those calculated for the remainder of the period.[31]

Reliance on the long-form sample produces a second measurement problem. The coverage of the establishment sample declines over time reflecting the reliance on long-forms in this study and their decreasing importance in terms of number of establishments over time. This should not greatly affect the rate of entry and exit when calculated as a proportion of number of firms or establishments at a point in time. The bias will be even less where entry and exit is measured in terms of employment, because of the relatively small size of the short-form establishments. Nevertheless, annual rates of entry using the long-form sample are calculated only for the period up to 1982. After that year, the sample does not have enough years at the moment to capture fully the transition of a short-form to a long-form plant. Therefore, it will increasingly underestimate entry rates. This can be overcome as more years of data become available.

- Variation in Census Coverage

The second problem arose because of a major change in census coverage. If left uncorrected, this change would have given a false increase in the entry reported in these two years and under-reported both entry and exit in prior years.

A major change in coverage in the Canadian Census of Manufactures occurred in 1978. In 1972, Statistics Canada lost a source of administrative information used to identify possible new establishments.[32] The result was a decline in coverage that was not rectified until 1978 and, to a lesser extent, 1979. In 1978, for example, 3,820 new establishments were added to the Census of Manufactures that Statistics Canada believed were already in existence. These "new" units accounted for 12 per cent of the total establishment count in 1978; however, since the majority were very small, and the increase in manufactured shipments due to their addition was much less significant, these "new" establishments accounted for only 1.7 per cent of the 1978 employment total.[33] In 1979, a further 1,142 preexisting establishments were added because of improvements in coverage. They accounted for only 3.3 per cent of the 1979 establishment total and only 0.37 per cent of the employment total.[34]

In order to correct for the change in coverage, the number of entrants and the employment associated with them that resulted from the increased coverage were identified and used to correct the entry and exit rates.

The correction employed for the 1978 and 1979 rates was straightforward. The overlap was subtracted.

The correction for previous years was more complicated. Because of the high death rate for new entrants, simple assignment of the 1978 and 1979 increased coverage figures to the earlier years would have understated earlier births. To correct for this, two assumptions were made: first, that the total number of births missed was distributed across the years 1972 to 1977 in proportion to those actually reported;[35] second, that the missed entrants died at the same rate after birth as those greenfield entrants actually reported. This allowed estimation of the missing entrants by year between 1972 and 1977. The employment associated therewith was calculated by assuming that the number of employees in each missed birth was the same as the average in those actually captured.

The exit rate data were also revised to allow for the fact that the undercoverage of entry in the mid 1970s would have led to a downward bias in calculated exit rates. Once again, the data for the rate of exit of greenfield entrants was used and applied to the additional entrants. The corrections have little effect on the average rate of entry or exit calculated over the decade.[36]

The Annual Enterprise Data Base

An enterprise is defined as all establishments in manufacturing and primary industries[37] under common control. If more of the enterprises's activity[38] is classified to a 4-digit manufacturing industry than any 4-digit industry in mining or logging, then the enterprise is classified to the manufacturing sector. Our sample of enterprises used for the short-run data base consists of those classified to the manufacturing sector.[39]

- Choice of sample

In the previous discussion, three reasons were adduced for excluding short-form establishments. These arguments also apply for enterprises that own short-form establishments. Such enterprises will tend to be almost exclusively single establishment enterprises, since establishments belonging to multi-industry, multi-establishment enterprises always complete long-form questionnaires.[40] Establishments that belong to single industry, multi-establishment enterprises are also likely to complete a long-form questionnaire, since they are large compared to single establishment enterprises.[41] In view of these factors, it was decided to exclude enterprises that (a) always owned only a single establishment (using the multi/single establishment code), and (b) the establishment always completed a short-form questionnaire.

The sample of enterprises thus consists of those classified to the manufacturing sector, but excludes those enterprises that always owned a single establishment that in turn always completed a short-form questionnaire for the Annual Census of Manufactures.

- Treatment of temporary exits

In a small number of instances, a plant or all of the establishments owned by an enterprise failed to report for a given year, but reported prior and subsequent to that year. If the rules outlined above were used, this would have been classified as an exit and subsequent entry, rather than a continuing plant or enterprise. Officials at Statistics Canada suggested that such failure to report could be due to a number of factors: a fire, strike, major overhaul of equipment, or slack demand. These situations were reclassified and the plant or firm was counted as continuing rather than as an exit or entrant.

- Determination of entry and exit method

The definition of enterprise exit and entry in the previous section made no attempt to distinguish between alternative methods of entry and exit. As has already been described, an enterprise may exit the manufacturing sector either by closing all of the plants it owns; or it may do so by selling the plants it owns to another enterprise--by divestiture. Equally, an enterprise may enter the manufacturing sector de novo, by building a new plant, or by purchasing plants of existing enterprises--by acquisition. In the analysis of long-run enterprise entry and exit, these different methods of entry and exit were differentiated.[42] The same distinction is made for the short-run estimates.

In considering the method of entry in the short run, the following approach was used to determine if the firm entered by acquisition, as opposed to plant creation: if the entrant first filed an Annual Census of Manufactures questionnaire in a particular year and the establishment(s) it owned in that year existed in the previous year, then the firm was classified as having entered by acquisition; if the plants did not exist in the previous year, then the enterprise was classified as having entered by plant creation. The same approach was used to distinguish the method of exit: if the exiting firm last filed an Annual Census of Manufactures questionnaire in a particular year and the plants it owned in that year were still alive in the next or subsequent year (but under a different owner), then the firm was classified as exiting by divestiture; if the plants did not file an Annual Census of Manufactures form in the next subsequent year, then the firm was classified as exiting by closing plant.[43]

A potential problem may arise either if an enterprise enters by both acquiring plants and building new plants, or if an enterprise exits by both divestiture and by plant closing. This could be handled by counting the firm twice or by creating a new category--entry by both acquisition and plant opening. Alternately, this firm could be assigned to one or other category on the basis of the importance of plants created in comparison to plants acquired.

The implications of using the first approach can be ascertained from the data that were employed to measure long-run entry and exit. While some firms entered both by building new plant and by acquiring it, the overlap was relatively small. These data come from comparing firm status in 1970 and 1979, a period that spanned a full decade. The possibility that a firm could enter by one route and then expand by the other should be greater for a ten-year period than for the annual period adopted to measure short-run entry. Thus, there should be much less overlap between the two methods of entry in any study that relies upon annual data.

In view of this, it was decided that it would be appropriate to count an entrant as either entering by plant creation or by plant acquisition. Therefore, an enterprise entrant was assigned to one or other of the two entry categories on the basis of the employment in the plants created versus the plants acquired. In those cases where an enterprise was classified as multi-plant, care was exercised to make sure the appropriate choice had been made.

CONCLUSION

The first section of this paper demonstrated how longitudinal panel data on firms and plants could contribute to the debate about the importance of the entry and exit process. The entry and exit process is just one of the forces at work that determine the strength of the competitive process. A more complete picture of the dynamics of the competitive system also requires an analysis of the extent to which the growth and decline of incumbents causes firms to change their relative rankings. It is also important to know whether mergers provide an important source of turnover. All these questions focus, in one way or another, on the extent and source of firm turnover. Equally important are a set of questions that focus on the effect of turnover on productivity and profitability.

Answers to all these questions require data bases that can track firms over time. The construction of these bases is not straightforward. Difficult problems have to be resolved. This paper has been devoted to a description of how they were met in the case of the Canadian data reported herein. Hopefully, they will serve as a source of reference for other researchers who are engaged in constructing similar data sets. Equally, they will permit the reader of the studies that are based on the data bases reported herein to evaluate both the strength and weakness of the research and to compare it, when appropriate, to the results of studies for other countries that use other data sources.

NOTES

[1] Of course, a merger may turn around an otherwise declining plant and generate employment in this fashion. Longer-run job turnover studies might then examine the extent to which acquired plants grew relative to the rest of the population.

[2] As such, it excludes head offices and similar activities if they are located separately from the establishment or if they serve more than one establishment. For further details, see Statistics Canada (1979, pp. 11-15).

[3] An establishment may undertake a number of different activities. To be classified to the manufacturing sector, the preponderance of these activities (based on value-added) must be in manufacturing. The manufacturing sector is defined as Division 5 of the 1970 Standard Industrial Classification. For details, see Dominion Bureau of Statistics (1970, pp. 23-43).

[4] There are a number of differing reporting units under the Census of Manufactures, including head offices and other auxiliary facilities. Attention is paid here only to establishments. For further details, see Statistics Canada (1979, p. 10).

[5] See, for further details, Statistics Canada (1979, pp. 17-18; 1983, pp. 23-25).

[6] In order to determine whether one legal entity controls another, attention is paid not only to cases where, directly or indirectly, one company "has more than 50 per cent of the exercisable voting rights of the subsidiary corporation"(Statistics Canada, 1979, p.17), but also to cases of minority control, "if factual information exists or acknowledgement by the entity in question is obtained" (Statistics Canada, 1983, p. 25).

[7] In some instances, several establishments may file a combined record. In these cases, the original statistics are projected by Statistics Canada across the individual establishment, each of which has a separate RSN.

[8] See McVey (1981, p. 72).

[9] The longitudinal enterprise code was maintained for the purposes of estimating concentration and foreign ownership statistics by J. McVey with the aid of J. Bousfield, B. Mersereau, and J. Lacroix.

[10] The results indicate that there was little difference in the annual entry and exit rates calculated with and without these exclusion criteria.

[11] Statistics Canada (1983) and "A Summary of the Establishment Description Tape File," Statistics Canada, unpublished internal working document, Ottawa, Appendix C-1. p 2.

[12] This is often the case for data bases used for U.S. studies that are generated from unemployment insurance or Dun and Bradstreet records. For a discussion of the problems with these data bases, see Baldwin and Gorecki (1990)

[13] The country of control categories were Canada, U.S., U.K., other Europe, and other foreign.

[14] A continuing firm is one that can be found in some 4-digit manufacturing industry in both the terminal and initial years of the comparison.

[15] A plant that is assigned to the entry by acquisition or exit by divestiture categories may also be involved in a horizontal merger. Such a merger may take place before or after the acquisition.

[16] See Johnson and Storey (1985) for a criticism of the Dun and Bradstreet data bases.

[17] See Storey (1985).

[18] The data for small plants that are taken from taxation administrative records in place of a mailed short-form questionnaire and the short-form records are both referred to here, for convenience, as "short-form".

[19] See Statistics Canada (1979, p. 44 and 1984, p. xiv). These figures refer to "small" establishments, which appear to be largely short-form establishments. See Statistics Canada (1979, pp. 43-44).

[20] These figures concerning short-form establishments for 1970, 1979, and 1982 are drawn from the same sources as footnote 19.

[21] Statistics Canada (1979, p. 42).

[22] To cite an earlier study on exit/entry conducted using census of manufactures data that excluded short-forms (McVey, 1981,p. 71).

[23] For a discussion of the total activity concept used in the census of manufactures, see Statistics Canada (1979, pp. 21-22). Measures of employment size for enterprises cover all employment including headquarters--that is, the employment of ancilliary units as well as that of operating establishments is included in the total.

[24] See Statistics Canada (1988) where this assumption is employed.

[25] Measures based on total employment are very similar to those based on shipments.

[26] See Baldwin and Gorecki (1983, Table 3, p. 15).

[27] See Baldwin and Gorecki (1983, Table 3, p. 15).

[28] Statistics Canada (1979, pp. 12-13).

[29] Statistics Canada, (1979, pp. 43-44).

[30] The amount of follow-up by Statistics Canada which determines whether an establishment should be classified as a long-form also varies over time. This will have less of an effect on this measure as long as an establishment that becomes large enough to receive a long-form is eventually caught. Of more concern is the probability that an entrant that is long-form upon entry is not caught by the system at the end of the year. The quality of the administrative data sources used and Statistics Canada's own reputation for diligence makes this unlikely.

[31] Although the cut-off point subsequently drifts upwards, the increase in the percentage of short-forms by 1983 is relatively minor--only about 4 percentage points. In light of the relatively small correction required for entry rates at the 1975 revision, which increased short-form establishments by 14 percentage points, the corrections were taken no further.

[32] Potter (1982, p. 21).

[33] Statistics Canada (1980, p. ix).

[34] Statistics Canada (1981, p. x).

[35] Alternate assumptions about the distribution of omitted entrants were found to have little impact on the mean of the annual birth and death rates for the decade.

[36] More detail can be found in Baldwin and Gorecki (1989h).

[37] In terms of the 1970 Standard Industrial Classification, these are Division 2, Major Group 1, Logging; Division 4, Mining (except Crude Petroleum and Natural Gas Industry and Major Group 5); and Division 5, Manufacturing. For full details, see Dominion Bureau of Statistics (1970, p. 17). In 1980, value-added of enterprises classified to manufacturing was $66,472 million; to mining, $9,062 million; and to logging, $702 million (Statistics Canada, 1983, Text Table VII, p. 15).

38. Manufacturing value-added is used to classify the enterprise to a 4-digit SIC on the basis of the largest unconsolidated enterprise owned by the consolidated enterprise. For details of these two enterprise concepts, see Statistics Canada (1983, pp. 28-30).

39. Using this definition of enterprises, there were 30,160 manufacturing enterprises in 1980 (Statistics Canada, 1983, Text Table VII, p. 15); however, if a manufacturing enterprise is defined as consisting only of establishments classified to the manufacturing sector, then there would be 30,197 enterprises classified to the manufacturing sector (Statistics Canada, 1983, Text Table XIII, p. 21). Hence there were 37 enterprises classified to mining or logging with activities in manufacturing. For example, a mining firm could own a small smelter. Hence, in terms of numbers, it makes little difference how we define the universe of manufacturing firms.

40. Statistics Canada (1979, p. 43) and McVey (1981, p.71).

41. For details, see Statistics Canada (1983, Text Table VII, p. 15).

42. No corresponding problem arises for establishment entry or exit. An establishment that exits the manufacturing sector--fails to file an Annual Census of Manufactures questionnaire--is assumed to exist no longer. In the terminology used here, it has exited by closing. Similarly, establishment entry can only occur de novo--the building of a new plant.

43. An alternative to matching whether an establishment filed an Annual Census of Manufactures questionnaire in year t and t+1 to determine whether the enterprise exited via closing plant is to refer directly to question 1.3.2 in the Annual Census of Manufactures questionnaire, which asks "Did this establishment go out of business during the reporting year?", to which the answer had to be "Yes" or "No" (Statistics Canada, 1979, p. 79). Work conducted within the Business Micro-data Integration and Analysis group of Statistics Canada suggested that matching the establishment Annual Census of Manufactures questionnaire between year t and t+1 was more reliable than accepting the answer to question 1.3.2.

References

Baldwin, J.R. and Gorecki, P.K. 1983. Entry and Exit to the Canadian Manufacturing Sector:1970-1979. Economic Council of Canada. Discussion Paper #225, Ottawa, February 1983.

Baldwin, J.R. and P.K. Gorecki. 1987. "Plant Creation Versus Plant Acquisition," International Journal of Industrial Organization 5: 25-41.

Baldwin, J.R. and P.K. Gorecki. 1989a. "Firm Entry and Exit in the Canadian Manufacturing Sector," Research Paper No. 23. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989b. "Intra-Industry Mobility in the Canadian Manufacturing Sector," Research Paper. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989c. "Mobility versus Concentration Statistics," Research Paper. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989d. "Productivity Growth and the Competitive Process: the role of firm and plant turnover," Research Paper. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989e. "Mergers Placed in the Context of Firm Turnover," Research Paper Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989f. "Job Turnover in Canadian Manufacturing," Research Paper No. 22. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989g. "Dimensions of Labour Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover," Research Paper No. 25. Analytical Studies Branch. Statistics Canada.

Baldwin, J.R. and P.K. Gorecki. 1989h. "Measuring Entry and Exit in Canadian Manufacturing Industry: Methodology," Research Paper. Analytical Studies Branch. Statistics Canada

Baldwin, J.R. and P.K. Gorecki. 1990. Structural Change and the Adjustment Process: Perspectives on Firm Growth and Worker Turnover. Economic Council of Canada.

Birch, D. 1979. The Job Generation Process. Cambridge(Mass): Massachusetts Institute of Technology Program on Neighborhood Change.

Birch, D. 1981. "Who Creates Jobs." The Public Interest. No.65, Fall, pp. 3-14.

Dominion Bureau of Statistics. 1970. Standard Industrial Classification,1970 . Cat. No. 12-501. Ottawa: Supply and Services Canada.

Johnson, S. and D. Storey. 1985. "Job Generation -- An International Survey: U.S. and Canadian Job Generation Studies Using Dun and Bradstreet Data: Some Methodological Issues," Research Working Paper No. 1. University of Newcastle-upon-Tyne.

McVey, J.S. 1981. Mergers, Plant Openings and Closings of large Transnational and other Enterprises;1970-1976. Cat. No. 67-507. Ottawa: Statistics Canada.

Potter, H.D, 1982. "The Census of Manufactures and the Labour Force Survey: some Experimental Approaches to Comparing Establishment and Household Survey Data," an occasional paper of the Analysis and Development Section of the Manufacturing and Primary Industries Division. Ottawa: Minister of Supply and Services.

Shepherd, W. 1984. "Contestability versus Competition," American Economic Review 74: 572-87.

Statistics Canada. 1979. Concepts and Definitions of the Census of Manufactures. Cat No. 31-528. Ottawa: Ministry of Industry Trade and Commerce.

Statistics Canada. 1980. Manufacturing Industries of Canada: National and Provincial Areas 1978. Cat. No. 31-203. Ottawa: Supply and Services.

Statistics Canada. 1981. Manufacturing Industries of Canada: National and Provincial Areas 1979. Cat. No. 31-203. Ottawa: Supply and Services.

Statistics Canada. 1983. Industrial Organization and Concentration in the Manufacturing, Mining, and Logging Industries, 1980. Cat. No. 31-402. Ottawa: Supply and Services Canada.

Statistics Canada. 1984. Manufacturing Industries of Canada: National and Provincial Areas 1982. Cat. No. 31-203. Ottawa: Supply and Services.

Statistics Canada. 1988. Building a Longitudinal Database of Firms in the Canadian Economy: The Case of Employment Dynamics. Ottawa: Minister of Supply and Services.

Storey, D.J. ed. 1985. Small Firms in Regional Economic Development: Britain, Ireland, and the United States. Cambridge: Cambridge University Press.

United States Small Business Administration. 1984. "The annual report on small business and competition." In Ronald Regan. The State of Small Business: A Report of the President. Washington (D.C.): U.S. Government Printing Office.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

RESIDUAL ERRORS AS THE MISSING LINK IN THE CYCLICAL ANALYSIS OF ECONOMETRIC MODELS AND THEIR DATA

L. Sager, P. Lin and T. Petersen*

ABSTRACT

For the past twenty years, the evaluation of estimated macroeconomic models has proceeded by critically comparing the non-stochastic simulated output from a number of simulation experiments. These experiments were carefully designed to expose the stability and cyclical properties of the models. By way of example, this paper argues in favour of a more balanced evaluation of econometric models using both the simulation technique and analytical techniques based on the model's eigenvalues, as suggested by Philip Howrey in the early 1970's. The example is an altered version of a recent macroeconomic simulation model constructed by Duguay and Rabeau (1987, 1988). The use of analytical techniques to evaluate econometric models has been greatly facilitated by the development of the LIMO package as part of the TROLL (1983) software system. We illustrate the use of the package to identify the various cyclical modes of a model and use its gain function to judge their relative importance. We then associate the cyclical modes with the model's structural parameters and with other parameters that reflect the autocorrelation of the residuals in the estimated equations.

1. INTRODUCTION

This paper uses an analytical technique based on an econometric model's eigenvalues to clarify the role played by the residuals of estimated equations in identifying its stability and cyclical properties. Placed in its historical context the paper is associated on the one hand, with Green, *et al* (1972), Evans, *et al* (1972) and Fromm, *et al* (1972) who used simulation techniques, and on the other hand with Howrey (1972) who used analytical techniques. To the extent that it is concerned with econometric model evaluation and comparisons, it is also loosely linked to the papers on the predictive performance of econometric models (Burmeister and Klein (1974)) and the cross-model comparison exercises of de Bever (1979) and O'Reilly (1983). For the most part these evaluation and comparison studies used simulation techniques.

In examining the stability and cyclical properties of econometric models, most analysts, with exceptions such as Howrey (1972) and Bergstrom (1984), have pursued a conventional approach by focusing their attention on impact and dynamic multipliers. The effect of an impulse shock on the model's endogenous variables is measured by simulating the model over time. For each endogenous variable, a dynamic multiplier is calculated at sequential points in time, first by taking the difference between the shocked solution and the control solution and then expressing this difference as a percentage of the size of the impulse shock in the exogenous variable. An examination of the dynamic multipliers over time reveals whether the model is stable (its multipliers becoming smaller), whether it exhibits cycles (the multipliers cycle over time) and what the periodicities of these cycles are.

Often the impulse multipliers tend not to cycle or display only short or long cycles that in many instances are quite weak. The long cycles tend to be ill-defined because the multipliers cannot be calculated over a long enough period of time. In any case the multipliers rarely reveal the richness of both long and short cycles that are found in the data used to construct the models.

Since the impulse multipliers failed to reveal both the long and short cycles that were found in the data underlying their models, analysts began to pay more attention to the nature of the shock. Although the impulse shocks were felt to be representative of the residual errors of the estimated behavioural equations of the model, in fact the residual errors displayed both auto and cross-correlation properties. These properties were not adequately reflected in the types of shocks that were imposed on the models. Thus attention was directed toward generating autocorrelated random shocks that reproduced the statistical properties of the sample-period residuals of the model's estimated equations. While the multipliers that resulted from the use

* L. Sager, P. Lin & T. Petersen, Industry Measures and Analysis Division, Statistics Canada, Ottawa, Ontario K1A 0T6. Discussions with D. Rhoades that clarified some of the issues in this paper are appreciated. Thanks are due to V. Campbell for preparing the manuscript.

of autocorrelated random disturbances showed both long and short cycles, these cycles tended to be both weak and heavily damped. This characterization of the cycles was obtained not only by observing the behaviour of the multipliers but also by estimating the spectra of the endogenous variables simulated over a long period of time. This was done for several of the major variables in the models of Green, Evans and Fromm and for several different realizations of the autocorrelated random disturbances.

In this paper we demonstrate using a representative econometric model, the advantages of applying analytical techniques rather than simulation techniques. These techniques are based on the calculation of the characteristic roots of the model and on the sensitivity of these roots to the model's parameters. With the aid of these techniques the stability and cyclical properties of the model are identified directly. More generally from the methodological point of view, these techniques serve as a powerful tool for the analysis of data in time and the models which are constructed from them, not only in economics, but also in many of the other social science disciplines.

2. THE MODEL

The model used for this analysis is patterned after that of a simulation model developed by Duguay and Rabeau (1987, 1988). Since the model is discussed in detail by them, only the way in which we altered it needs to be discussed. As structured by Duguay and Rabeau, the model reflects a closed economy. More importantly, the coefficient values are imposed rather than estimated. They are chosen with the intention that they should reflect empirical evidence about the structure of the economy. The model is structured so that the coefficients can be easily changed to switch its emphasis from having a short-run Keynesian outlook to having a long-run neoclassical steady-state growth outlook. The model contains five main blocks: private demand, output and employment, prices and wages, a government sector and a monetary and financial sector.

The model used for the analysis in this paper departs from the Duguay and Rabeau structure in two ways. First it reflects an open economy by including equations for the imports and exports of goods and services and the exchange rate. Imports and exports are simple functions of domestic and foreign incomes and relative exchange-rate adjusted prices. The change in the exchange rate is a function of changes in the trade-balance-to-real-income ratio, changes in the spread between domestic and foreign interest rates and the change in government debt held by foreigners. This formulation, a major departure from the Duguay-Rabeau framework, permits one other channel, the sale of bonds to foreigners, by which the government might offset an increase in real interest payments. Government bonds held by foreigners are a function of these real interest payments.

The second way in which the model differs from that of Duguay and Rabeau is that the coefficient values, rather than being imposed, are estimated using for the most part, quarterly, seasonally adjusted National Accounts data. The period of estimation is 1966 to 1988. This approach allows the coefficients to be estimated and the regression residuals to be calculated.

There are 58 endogenous variables in the model; thirty-five of these are identities and 23 are estimated by regression methods.[1]

3. METHODOLOGY

This study applies <u>L</u>inear <u>Mo</u>del Analysis (LIMO) to the model outlined above to show the stability and cyclical properties of the model and the effect on these, when the parameter estimates are corrected for the autocorrelation in the residuals of the stochastic equations. A brief description of the theory underlying LIMO follows.

A general stochastic structural form model can be described as

$$(1) \quad f((y_{g.t},\ldots,y_{g.t-m_g}\},(x_{k.t},\ldots,x_{k.t-n_k}),\{\beta_j\}) = (e_{g.t})$$

where $g = 1,\ldots,G$, $k = 1,\ldots,K$, and $j = 1,\ldots,J$,
f is a vector of G functional relations,
$y_{g.t}$ is one of G endogenous variables,
$x_{k.t}$ is one of K exogenous variables,
β_j is one of J constant coefficients,
$e_{g.t}$ is one of G residuals,
m_g is the maximum lag in y_g appearing in any of the G functional relations,

n_k is the maximum lag in x_k appearing in any of the G functional relations.

Assuming that the residuals take their expected value of zero, the deterministic structural form model can be manipulated to arrive at (2), the linearized state-space representation of the structural form model.

$$(2) \quad D_t.\Delta z_t = E_t.\Delta z_{t-1} + F_t.\Delta x_t + H_t.\Delta\beta$$

where Δz_t, Δx_t and $\Delta\beta$ are deviations of the state-space variables and parameters around a simulation path. D_t, E_t, F_t and H_t are matrices of derivatives of the state-space representation of the structural form model with respect to current period endogenous variables, lagged endogenous variables, exogenous variables and parameters respectively.

Assuming D is invertible the linearized state-space representation of the reduced form model is given by (3):

$$\cdot\ (3) \quad \Delta z_t = A_t.\Delta z_{t-1} + B_t.\Delta x_t + C_t.\Delta\beta$$

As is evident from equations (2) and (3) LIMO will only deal with systems of deterministic equations for which the residuals of the stochastic equations are assumed to be zero. An alternative procedure is needed for systems of equations with autocorrelated disturbances or these equations must be transformed in such a way as to make them compatible with the LIMO environment.

Assume that the residuals e_g, of one of the G stochastic endogenous variables (y_g) of (1) are autocorrelated:

$$(4) \quad y_g = f_g + e_g$$

If u_t is an independent random variable and P(L) a polynomial in the lag operator L then the autocorrelation in e_g can be expressed by:

$$e_{g.t} = p_1 e_{g.t-1} + p_2 e_{g.t-2} + \ldots + u_t \quad \text{or} \quad e_{g.t} = \frac{u_t}{(1-P(L))}$$

Transforming (4) by multiplying through by the polynomial (1-P(L)) and normalizing on y_g leads to:

$$(5) \quad y_g = P(L)y_g + (1-P(L))f_g + u_t$$

The implications of (5) for the analysis of a model by LIMO as reflected in (1) to (3) are straightforward. Additional lags of the endogenous variable will be introduced in (1), the coefficient vector β must be expanded to reflect the coefficients of the polynomial (1-P(L)), and additional lags for the exogenous variables in f_g must be added. When equations with autocorrelated errors are viewed in this way, they can be included in a model and in a LIMO analysis of the model's stability and cyclical structure.[2]

The matrix A of (3) is called the dynamics matrix and it is the only link between the behaviour of the linearized model in past periods and its behaviour in the current period. While this behaviour depends on the linearized version of the model, Kuh *et al* (1985, p. 14) stress the benefits of using the analytical method outlined above over other nonlinear simulation methods. The model's behaviour is reflected in the characteristic roots of A. The root magnitudes reveal the stability of the model and their imaginary parts provide information about its cyclical properties. Root magnitudes greater than one indicate that the model is unstable. Once perturbed from its initial state by an impulse shock the model will not return to its original state. On the other hand, a model whose largest root is less than one will return to its initial state, but more slowly than a model whose largest root magnitude is smaller. The imaginary parts of the roots are useful in describing the cycles in the model. The amplitude of these cycles can be growing, constant or damped depending on whether the magnitude of the complex root is greater than, equal to or less than unity. Cycles are measured in terms of periodicity which is the length of a complete cycle in the units of time (months, quarters, years) inherent in the model.[3]

4. DYNAMIC PROPERTIES BASED ON CHARACTERISTIC ROOTS

To reveal the effect of accounting for autocorrelation in the residuals on the model's stability and cyclical properties, three versions of the model are analyzed. Ordinary least squares is used to estimate Version 1 with no account being taken of the autocorrelation in the estimated residuals. Generalized least squares is used to estimate Version 2. The residual errors are assumed to follow a second-order autoregressive process. A process of this order is needed to generate the cycles observed in the data for the left-hand-side stochastic variables. For Version 2, the more efficiently estimated structural coefficients are used to construct the A matrix. In other words Version 2 incorporates the effects of modelling the error term on the estimates of the structural parameters but does not incorporate the coefficients of the error structure themselves in the dynamic matrix A. This version is sometimes referred to as the "AR=0 version". Version 3 is the same as Version 2 except that the coefficients of the error structure are included in the A matrix by using equation (5). This version is sometimes referred to as the "AR≠0 version".[4]

Table 1 presents selected information about the characteristic roots for these three different versions of the model. The criteria for judging this root information is how well it represents the cycles that are found in the data for the stochastic endogenous variables of the model. The cycles found in these data are summarized in the histogram Figure 1, Panel 1.

The histogram in Figure 1, Panel 1 shows the distribution by cycle length, of the statistically significant peaks in the periodograms of the stochastic endogenous variables.[5] Whether the ordinate at a given frequency in the periodogram for a given variable differs from those at other frequencies is tested using Fisher's g-statistic. (See Priestley (1981), pp. 406-412.) At the 0.1 significance level for this test, the null hypothesis that the ordinates come from a white noise process was rejected in 41 cases. Of these, 15 represent long cycles between 11 and 22 years, 12 represent short cycles between 2 and 3 years long and the remainder are fairly evenly split between cycles ranging from 3 to 11 years long.

Another way to characterize the cyclical behaviour of the data for the stochastic endogenous variables is to calculate their average periodogram. This periodogram, shown as the solid line in Figure 2, is the average of the ordinates of the periodograms for each of the individual variables. Before averaging, the ordinates of each periodogram were normalized by dividing by the variance of the detrended series. This normalization served two purposes. The first was to make the periodograms more comparable across the variables. The second was to make the average periodogram more comparable to the histogram of statistically significant peaks shown in Figure 1, Panel 1. After averaging, the ordinates were again normalized by dividing by the largest ordinate value. The purpose of this normalization was to make the average periodogram more comparable to the gain functions that are analyzed in Section 5 for various versions of the model (see also Footnote 6).

Comparing the average periodogram with the histogram shown in Figure 1, Panel 1, it can be seen that the direction of change indicating the importance of cycles at various periodicities is the same in eight out of ten cases. Movement in the opposite direction occurs between cycles 7.3 and 5.5 years long and between cycles 3.1 and 2.7 years long. In addition, the ordinates of the average periodogram suggest that short cycles are relatively less important than long ones... a view that is missed by looking at the histogram alone.

By comparing the cyclical information for the stochastic endogenous variable data shown in Figure 1, Panel 1 with the characteristic root information for Version 1 of the model shown in Table 1, several things can be noted. The largest complex root magnitude is 1.03 and is associated with an explosive cycle whose periodicity is 17.2 years. Roots 10, 11 and 12 are less than one and imply that their associated cycles which vary between 14 and 16 years, are damped. These cycles reflect the long 11-to-22-year cycles in Figure 1, Panel 1. However the most important point to note about the comparison is that according to the root information, Version 1 does not reflect the short 2-to-3-year cycles that were shown to be in the stochastic endogenous data. Since these short cycles appear in the data but not in the characteristic roots of the A matrix that is derived from the estimated coefficients of Version 1, they must reside in the estimated residuals of the model.

As indicated above, accounting for the autocorrelation in the residuals of the estimated equations led to Version 2 of the model. For this version, Table 1 shows that while still unstable, its largest real root has fallen from 1.13 to 1.02. The largest complex root magnitude is associated with a 23.4-year cycle. The complex roots with magnitudes less than unity imply damped cycles whose lengths are greater than 14 years. While root 15 indicates the presence of a 3-year cycle, its small root magnitude implies that the amplitude of this cycle is small relative to those of the other cycles in the model. Aside from this extremely small root, there are no other complex roots associated with short-run cycles. This indicates that although the coefficients of the AR=0 version may be more efficiently estimated, when collected together into the A matrix, they still do not give a version of the model that reflects the short cycles found in the data.

To demonstrate that short cycles remain in the residuals (e_q), of the estimated equations, these residuals were calculated for the normalized forms of (5) that would be used to solve the model in a simulation context but with the autoregressive polynomial coefficients set to zero. It should be noted that in many cases these normalized equations are transformations of the estimated equations. For some equations, constraints have been imposed on the estimated coefficients for reasons of economic theory. This implies that the dependent variable in estimation is a combination of the normalized dependent variable and one or more independent variables. Another example might be the estimation of an equation in logarithmic terms while its normalized form requires it to be converted to levels by taking antilogs. Thus in general, the properties of the residuals of the normalized equations may be quite different from those of the estimated equation.

The histogram in Figure 1, Panel 2 shows the distribution by cycle length, of the statistically significant peaks in the periodograms of the residuals of the normalized equations for the stochastic endogenous variables. At the 0.1 level of significance, the null hypothesis that the ordinates of these periodograms come from white noise processes was rejected for 40 cases. Of these, 19 represent long cycles of between 11 and 22 years, 6 represent short cycles between 2 and 3 years and 10 represent cycles whose length ranges between 5 and 11 years. A comparison of the cyclical information contained in this histogram with the root information for Version 2, Table 1, shows that the AR=0 version of the model still leaves much of the explanation of the short cycle in the normalized residuals of the model's stochastic equations.

While Version 2 of the model uses the more efficiently estimated parameters that result from implementing some form of the generalized least squares estimator in the estimation process, the transformations implied by (5) for the endogenous and exogenous variables are not reflected in the normalized equations used to solve the model in the simulation context. Since the normalized equations do not reflect the transformations, the residuals calculated from them and used to construct Figure 1, Panel 2 do not do so either. Accounting for these transformations implies the incorporation of additional lags in the endogenous and exogenous variables in the normalized equations of the model. The effect of this is that the influence of the estimated autocorrelation coefficients for the error processes will be reflected in the A matrix only for the AR≠0 version of the model.

The characteristic root information for Version 3, Table 1, shows that the largest root magnitude increases from 1.02 (Version 2) to 1.07. More important, the largest complex root is associated with an explosive cycle whose periodicity is slightly greater than 2 years. Roots 12, 15, 16 and 25, all with magnitudes greater than .87 are associated with stable cycles whose periodicity ranges from 16 to 22 years. Root 18 with a magnitude of 0.95 reflects a cycle whose periodicity is 8.5 years. These roots imply that the cyclical structure of Version 3 corresponds more closely with the cyclical structure of the data shown in Figure 1, Panel 1.

Since much of the cyclical structure that is in the data for the stochastic endogenous variables is reflected in the root structure of Version 3, this implies that there should be little of it left in the residuals for the normalized forms of (5) used to solve the model in a simulation context. The histogram given in Figure 1, Panel 3 shows the distribution of the statistically significant peaks in the periodograms for these residuals. At the 0.1 level of significance, the null hypothesis that the ordinates of these periodograms come from white noise processes can be rejected for only 5 cases. Of these, three occur for cycles 22 years long, one for a 5-year cycle and one for a 7-year cycle. The histogram reflects no significant cycles in the 2-to-3-year range.

The cyclical behaviour of the model as portrayed by its characteristic roots looks more like the cyclical structure of the data for the stochastic endogenous variables when the coefficients reflecting the autoregressive models for the error processes appear in the A matrix.

5. THE GAIN FUNCTION, VERSION 3

This section examines the importance of the 2-to-3-year cycle relative to cycles of other lengths in the overall cyclical structure of Version 3. In isolation, the root magnitude for any particular cycle provides little information about its contribution to the overall cyclical structure of the model. For any root, the larger its magnitude the stronger will be its associated cycle. However for models with several roots it is difficult to infer the strength of a particular cyclical mode in the model from the magnitudes of the individual roots. This is because both the periodicity spacing between roots and the number of roots near a given periodicity have a bearing on the amplitude of any particular cyclical mode in the model.

The analytical technique used to determine the importance of the 2-to-3-year cycle in the overall cyclical structure of Version 3 is the gain function.[6] The gain function combines and transforms the information pertaining to all the cyclical roots in such a way as to make clear through relative amplitudes, the importance of any particular cyclical mode in the model. The gain function shows the squared amplitude of the cycles in the endogenous variables in the model as a function of periodicities. Since only relative amplitudes are of interest, the gain functions have been normalized so that the largest ordinate has a value of 1.

The gain function for the major cyclical roots (3, 12, 15, 16, 18, 25 and 34) for Version 3 as identified in Table 1, is shown as the long dashed line in Figure 2. The gain function has its maximum amplitude at the 22-year cycle. It falls close to zero at the next observable periodicity which corresponds to the 11-year cycle. More important the amplitude of the gain function at the 2-year cycle is near zero. This shape for the gain function implies that the model is dominated by cycles with long trend-like periodicities. While the characteristic root information for Version 3 implied the existence of a 2-year cycle (with a large root magnitude), the gain function which places this 2-year cycle in the context of all other cycles in the model, shows it to be very weak.

As an aside but as a matter of interest, the gain function corresponding to the characteristic root (Root index 3) that reflects the 2-year cycle is shown as the short dashed line in Figure 2. The gain function for this root alone has an amplitude that is near zero at the 22-year cycle and that rises to its maximum value at the 2-year cycle. This gain function taken by itself confirms the existence of the 2-year cycle in the model. However in the context of the overall cyclical structure of the model, this 2-year cycle is very weak.

Another point illustrated in Figure 2 is that the average periodogram for the stochastic endogenous data (the solid line in this figure) has a shape that is in general, similar to that of the gain function that reflects the complex roots of Version 3 of the model. With the exception of its minor upward movements at the 3.7-year and the 2.4-year cycle, it slopes downward continuously from left to right like the gain function. However its rate of descent is not as steep.

6. CHARACTERISTIC ROOT SENSITIVITIES

This section focuses attention on those roots of Version 3 that generate significant cycles in the 2-to-20-year range. The objective is to determine which parts of the model's structure contribute most to these roots. To accomplish this, the sensitivity of the roots to small perturbations in the structural parameters is determined. Root sensitivities can be calculated for the root period and the root magnitude. The root period sensitivity[7] in elasticity form, shows the percentage change in periodicity in response to a one percent change in the i,jth element of the D or E matrix. The root period sensitivities calculated from the D matrix provide information about the relationship between a particular root period and the parameters of current endogenous variables. The relationship between the root period and the parameters of lagged endogenous variables can be observed through the root period sensitivities calculated with respect to entries in the E matrix. As with the period sensitivities, root magnitude sensitivities can be calculated with respect to the D or E matrix. The root magnitude sensitivity in elasticity form, shows the percent change in the magnitude of a root in response to a one percent change in the i,jth element of the D or E matrix.

Table 2 summarizes selected root period sensitivities for Version 3. Each row reflects variables appearing in a particular equation in the model. Columns 1 and 2 refer to the current left-hand-side endogenous variable for a particular equation. The other columns refer to lagged endogenous variables (whether for the same equation or other equations) that are part of the right-hand-side determinants of this equation. For example in the first row, the current endogenous variable, KAP, is the left-hand-side variable of the equation for KAP and the lagged variables KAP(-2) and KAP(-3) are two of the right-hand-side determinants of this equation. Based on the root period sensitivities, Root 3 reflecting the 2-year cycle, is closely associated with the capital stock (KAP) and the short-term nominal interest rate (RS). A one percent increase in the structural coefficient associated with RS causes the periodicity of Root 3 to increase by 3.87 per cent. Similarly, the root period can be reduced by 9.39 per cent as the result of a one percent increase in the structural coefficient associated with RS(-1) in the equation for RS.

Root 12, reflecting the 20-year cycle, is mainly related to the expected rate of inflation (DNPE) and its lagged values.

Roots 15 and 16, both reflecting the 16-year cycle, are dominated by the trend rate of growth of employment (DNNE) and the trend rate of growth of total factor productivity (DNTFPE). The large values indicate that the root periodicities are highly sensitive to the structural coefficients associated with current and lagged DNNE and DNTFPE in the equations for these variables.

Root 18, reflecting the 9-year cycle, is related to RS, DNPE, KAP, and private sector employment (NIC). Of these four variables, DNPE dominates the root-variable association. The root period is more sensitive to the coefficients associated with the lagged variables RS, DNPE and KAP than to the current variables.

Having noted that Root 12 is associated with DNPE and Roots 15 and 16 with DNNE and DNTFPE, it would be interesting to know how the stability and cyclical characteristics of Version 3 would change if these variables were exogenized. Exogenizing these variables by removing their associated equations and coefficients from Version 3, leads to Version 4. Selected root information for Version 4 is shown in Table 1. The stability of Version 4 in terms of its largest root is unchanged from Version 3. The 2-year and 9-year cycles are retained. However Version 3's 16- and 20-year cycles which the sensitivity analysis indicated were closely associated with the equations for DNNE, DNPE and DNTFPE, disappear in Version 4. This confirms the hypotheses pointed to by the sensitivity analysis. It illustrates that a characteristic root sensitivity analysis, combined with a careful analysis of the economic hypotheses on which the model is constructed, can lead to inferences about which elements of a model's structure are responsible for the cyclical properties of the model.

The gain function for the major cyclical roots 3, 14 and 29 as noted in Table I for Version 4 is shown as the dotted line in Figure 2. This gain function reaches its maximum amplitude at the 11-year cycle and falls to near zero at the 3.7-year cycle. However unlike the gain function for Version 3, it shows some amplitude (although small) in the 2-year cycle range.

7. SOME FINAL COMMENTS

From the methodological point of view, the observations concerning the stability and cyclical properties of a model as derived from the conventional simulation approach seem naive compared to those that can be made following the analytical approach. This approach based on the eigenvalues of the linearized model, showed Version 2 to be unstable with 14-to-23-year cycles. Version 3 was shown to be unstable and to possess long 16-to-22-year cycles and a 2-year cycle that was much weaker than the impression given by Panel 1 in Figure 1. Incorporating the autocorrelation of the estimated residuals in the dynamic matrix A, permitted the two-year cycle found in the data, to be reflected in the model's cyclical structure. While accounting for this 2-year cycle may be important for forecasting exercises, its existence depends on introducing additional lagged endogenous and exogenous variables explicitly. It is important to note that while root analysis pointed to the existence of the 2-year cycle, its weakness relative to that of the trend-like cycles was only established by calculating the model's gain function.

The root period sensitivity analysis illustrated the usefulness of the analytical approach for identifying those coefficients in a model's structure that are most closely associated with the appearance of particular cyclical modes in the model. While this analysis pointed to the coefficients associated with the capital stock and the short-term interest rate as being important determinants of the 2-year cycle, the coefficients associated with these variables appear in the equations of the model as the result of accounting for the autocorrelation processes of the residual errors. An economic explanation of the 2-year cycle in terms of the interaction between these two variables as reflected in their structural coefficients is not obvious. It may be that much of the residual autocorrelation that gave rise to the 2-year cycle is indicative of a misspecification of the economic hypotheses upon which the model is built.

FOOTNOTES

[1] The stochastic variables are Total consumption, Trend rate of growth of employment, Price expectations (measured as the trend rate of growth of P), Trend rate of growth of total factor productivity, Level of personal exemptions related to personal income taxes, Government bonds held by foreigners, Current government non-wage expenditures, Government transfers to persons, Supply of high-powered money, Private and government sector capital stock, Private sector stock of inventories, Imports of goods and services, Private sector employment, Aggregate price level, Foreign exchange rate, Risk premium on real capital, Risk premium on long-term bonds, Long-term

interest rate, Short-term interest rate, Average rate of corporate income tax, Average rate of personal income tax, Private sector wage rate, and Exports of goods and services.

2. In their simulations with the Brookings Model, Fromm *et al* (1972) found that transforming equations to reflect their autocorrelated errors gave poor results in complete system simulations. They argued that the transformed equation uses more lagged values for the endogenous variables and thus permits the simulated and actual values to deviate through error build-up in dynamic simulation. For the OBE Model (Green, *et al* (1972) and the Wharton Model (Evans, *et al* (1972)) autoregressive transformations for individual equations brought about modest improvements in the forecasting properties of the models over the sample period simulations. However the effect of these transformations on the cyclical properties of the models was not discussed.

3. If $R_h = a_h+b_h.i$ defines the h'th complex root where a and b are its real and imaginary parts respectively and $i=\sqrt{-1}$, then the periodicity of this complex root is defined as $2\pi/(\text{Arctan } (b_h/a_h))$.

4. In their long-run simulation studies, Fromm *et al* (1972), Green *et al* (1972) and Evans *et al* (1972) adjusted the constant terms of their equations by a factor which reflected the first-order autocorrelation of the equations' residuals rather than using transformed equations such as (5).

5. Given T observations $X_1,\ldots,X_T$, the function P(f) called the periodogram, is defined for all f (frequency) in the range $-\pi \leq f \leq \pi$ by

$$P(f) = \frac{1}{2\pi}\left|\sum_{t=0}^{T-1} X_t\ e^{-i2\pi ft}\right|^2$$

where P(f) is evaluated at the set of frequencies 0, $2\pi/t$, $4\pi/t,\ldots$. The construction of the periodogram assumes that the series is stationary. In order not to eliminate cycles whose periodicity might be very long, the variables were detrended by using a simple linear time trend.

6. Each of the characteristic roots of the model may be thought of as being related to a polynomial g(L) in the lag operator L in the time domain. The response of any endogenous variable Y_t, in the model to a random variable X_t is given by

$$\text{(a)}\quad Y_t = \sum_{-\infty}^{\infty} g(L)\ X(t-L)$$

The magnitude and periodicity of each of the characteristic roots provide information about the form of g(L). Fourier transforming (a) and taking expected values lead to (b), the spectral representation of the response function, called the gain function, in the frequency domain.

$$\text{(b)}\quad \Gamma_{YY}(f) = T(f).\Gamma_{XX}(f)$$

where Γ_{YY} and Γ_{XX} are spectral density functions and T(f) is defined by (c)

$$\text{(c)}\quad T(f) = \sum_{-\infty}^{\infty} g(L)\ e^{-ifL}$$

From (b) it can be seen that depending on whether T(f) is greater than, equal to or less than one, the amplitude of the cycle at any frequency f, in Y will be greater than, equal to or less than the amplitude of the cycle at the corresponding frequency in X. Priestley (1981, p. 273) shows that the gain function for the model is equal to the product of the gain functions corresponding to each of its characteristic roots.

7. The root period and magnitude sensitivities in elasticity form, can be calculated from the D or E matrix of the linearized system as

	Period	Magnitude
From D matrix	$\dfrac{\partial\lvert t_h\rvert}{\partial d_{ij}}\cdot\dfrac{d_{ij}}{\lvert t_h\rvert}$	$\dfrac{\partial\lvert R_h\rvert}{\partial d_{ij}}\cdot\dfrac{d_{ij}}{\lvert R_h\rvert}$
From E matrix	$\dfrac{\partial t_h}{\partial e_{ij}}\cdot\dfrac{e_{ij}}{\lvert t_h\rvert}$	$\dfrac{\partial R_h}{\partial e_{ij}}\cdot\dfrac{e_{ij}}{\lvert R_h\rvert}$

where R_h is the modulus of the h'th characteristic root; d_{ij} and e_{ij} are the i,j'th entries of the D and E matrices respectively; and t_h is the periodicity of the h'th characteristic root.

REFERENCES

Bergstrom, A.R., (1984), "Monetary, Fiscal and Exchange Rate Policy in a Continuous-Time Econometric Model of the United Kingdom" in *Contemporary Macroeconomic Modelling*, P. Malgrange and P. Muet (editors), London: Basil Blackwell.

Burmeister, E. and L. Klein, (1974), "Symposium Econometric Model Performance: Comparative Simulation Studies of Models of the U.S. Economy", *International Economic Review*, June 1974, Oct. 1974 and Feb. 1975.

deBever, L., D.K. Foot, J.F. Helliwell, G.V. Jump, T. Maxwell, J.A. Sawyer and H.E. Waslander, (1979), "Dynamic Properties of Four Canadian Macro-Models: a Collaborative Research Project", *Canadian Journal of Economics*, 12(2), 133-194.

Duguay, P. and Y. Rabeau, (1987), *Les Effets Macro-Economiques des Déficits Budgétaires: Resultats d'un Modele de Simulation*, Rapport Technique 47, Ottawa, Banque du Canada, novembre.

Duguay, P. and Y. Rabeau, (1988), "A Simulation Model of Macro-Economic Effects of Deficit", *Journal of Macroeconomics*, Vol. 10, Number 4, Fall, 539-564.

Evans, M.K., L.R. Klein and M. Saito, (1972), "Short-Run Prediction and Long-Run Simulation of the Wharton Model" in B. Hickman, *Econometric Models of Cyclical Behavior*, Vol. 1, 137-200.

Fromm, G., L.R. Klein and G. Schink, (1972), "Short- and Long-Term Simulations with the Brookings Model" in *Econometric Models of Cyclical Behavior*, Vol. 1, 201-310.

Green, G., M. Liebenberg and A. Hirsch, (1972), "Short- and Long-Term Simulations with the OBE Econometric Model" in *Econometric Models of Cyclical Behavior*, Vol. 1, 25-136.

Hickman, B. (editor), (1972), *Econometric Models of Cyclical Behavior, Vol. I and II*, New York: National Bureau of Economic Research, Columbia University Press.

Howrey, E.P., (1972), "Dynamic Properties of a Condensed Version of the Wharton Model" in *Econometric Models of Cyclical Behavior*, Vol. 2, 601-671.

Kuh, E., J.W. Neese and P. Hollinger, (1985), *Structural Sensitivity in Econometric Models*, New York: John Wiley & Sons.

O'Reilly, B., G. Paulin and P. Smith, (1983), *Responses of Various Econometric Models to Selected Economic Policy Shocks*, Technical Report No. 38, Ottawa: Bank of Canada.

Priestley, M.B., (1981), *Spectral Analysis of Time Series, Vol. 1: Univariate Series*, London: Academic Press.

Troll (1983), *Program LIMO*, Technical Report No. 34, Massachusetts Institute of Technology, Cambridge, Mass., December.

TABLE 1: SELECTED CHARACTERISTIC ROOT INFORMATION

	Version 1		Version 2		Version 3		Version 4	
Characteristic Roots (index)	Root Magnitude	Periodicity (years)	Root Magnitude	Periodicity (years)	Root Magnitude	Periodicity (years)	Root Magnitude	Periodicity (years)
1	1.13		1.02		1.07		1.07	
2	1.03		1.02	23.43	1.03		1.03	
3	1.03	17.2	1.01		1.01	2.08	1.01	2.02
4	1.02		1.01		1.01		1.01	
5	1.01		0.99		1.01		1.01	
6	1.00		0.99		1.00		1.00	
7	1.00	92.17	0.99	89.86	1.0		1.00	
8	1.00		0.96	15.88	1.0		1.00	
9	0.99		0.96	15.88	1.0	67.48	1.00	67.43
10	0.96	15.89	0.94	89.14	1.0		1.00	
11	0.96	15.88	0.92	14.24	1.0		1.00	
12	0.94	14.20	0.92		0.99	19.78	0.99	
13	0.93	55.56	0.92	180.56	0.99		0.97	
14	0.93		0.73	21.60	0.97		0.96	8.89
15	0.92		0.03	3.48	0.96	15.89	0.96	
16	0.92				0.96	15.89	0.94	
17	0.83				0.96		0.91	
18	0.00				0.95	8.51	0.86	
25					0.87	21.96		
29							0.58	11.94
34					0.58	11.94		

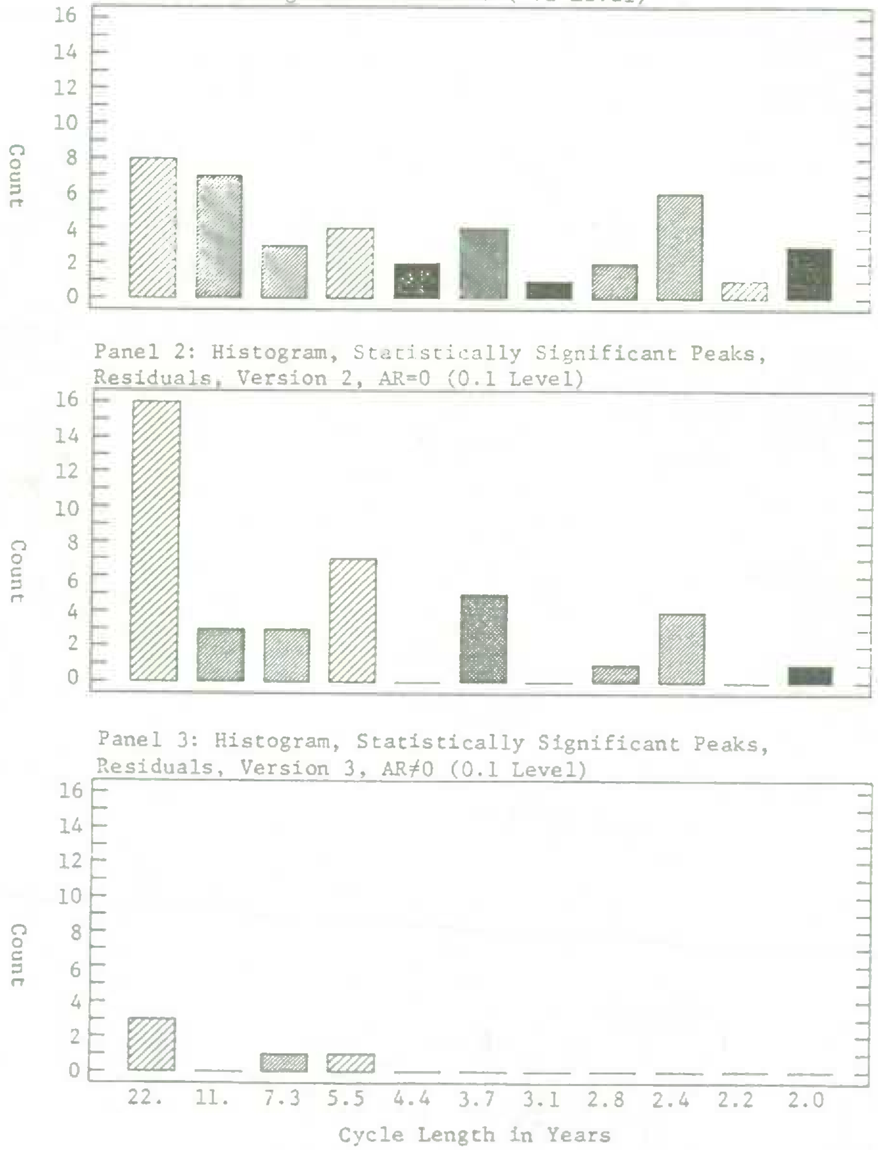
Note: Only characteristic roots that are complex give rise to cycles whose periodicity is measured in years (see Footnote 3).

TABLE 2: SELECTED ROOT PERIOD SENSITIVITIES, VERSION 3

(1) Variable	(2) Sensitivity	(3) Variable	(4) Sensitivity	(5) Variable	(6) Sensitivity	(7) Variable	(8) Sensitivity
			ROOT 3				
KAP	-1.33			KAP(-2)	2.70	KAP(-3)	-1.21
RS	3.87	RS(-1)	-9.39	RS(-2)	-2.95		
			ROOT 12				
DNPE	-444.27	DNPE(-1)	1229.07	DNPE(-2)	-1128.83	DNPE(-3)	343.91
			ROOT 15				
DNNE	133.26	DNNE(-1)	-486.09	DNNE(-2)	566.39	DNNE(-3)	-213.56
DNTFPE	-392.91	DN1FPE(-1)	1128.56	DNTFPE(-2)	-1072.38	DNTFPE(-3)	336.73
			ROOT 16				
DNNE	-392.92	DNNE(-1)	1128.59	DNNE(-2)	-1072.41	DNNE(-3)	336.74
DNTFPE	133.28	DNTFPE(-1)	-486.16	DNTFPE(-2)	566.45	DNTFPE(-3)	-213.58
			ROOT 18				
RS	-4.79	RS(-1)	7.63	RS(-2)	-2.80		
DNPE	-12.02	DNPE(-1)	34.20	DNPE(-2)	-31.33	DNPE(-3)	9.18
KAP	-4.30	KAP(-1)	11.13	KAP(-2)	-9.10	KAP(-3)	2.32
NIC	-1.12	NIC(-1)	0.89	NIC(-2)	0.87		

Note: For any given row, the variable in column (1) refers to the left-hand-side variable for the equation it represents in Version 3. Variables in columns (3), (5) and (7) are variables (primarily lagged endogenous variables) appearing as part of the right-hand side of the same equation. Only the largest root sensitivities are reported.

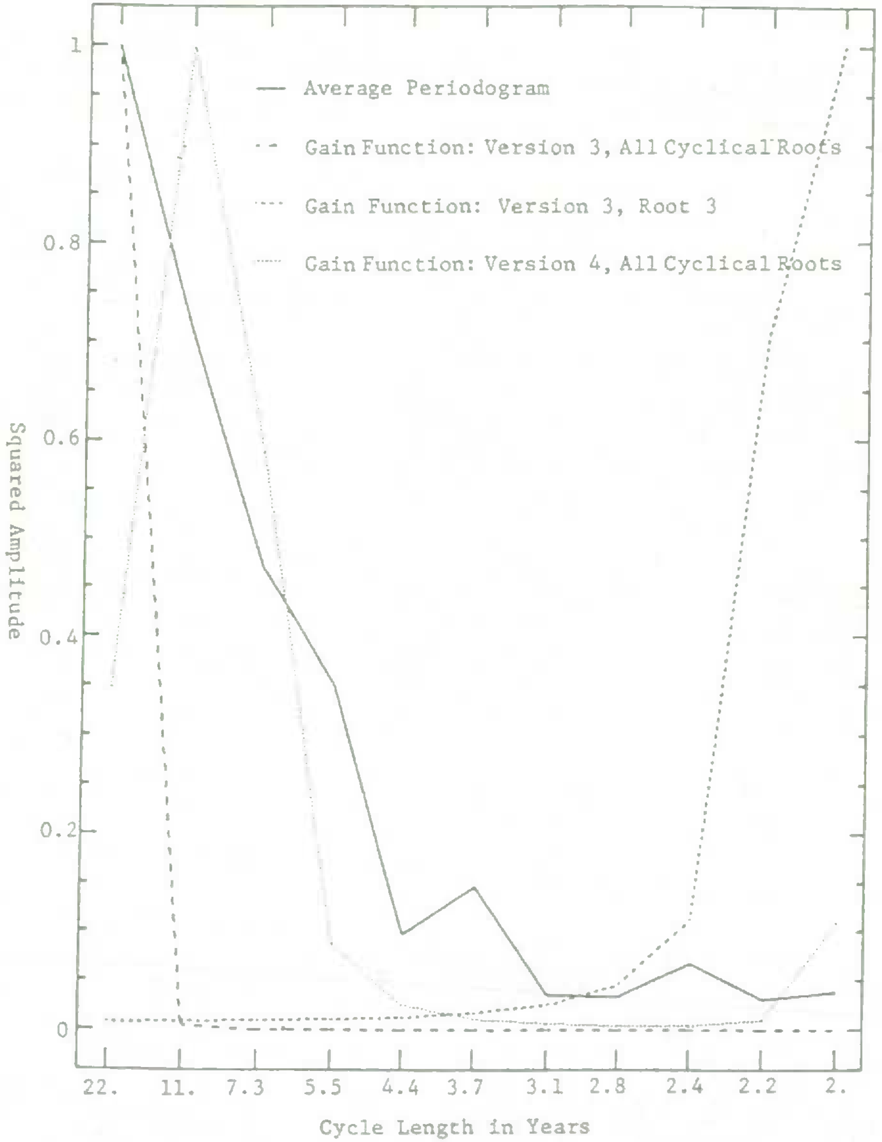
Figure 1: Panel 1: Histogram, Statistically Significant Peaks, Stochastic Endogenous Variables (0.1 Level)



Panel 2: Histogram, Statistically Significant Peaks, Residuals, Version 2, AR=0 (0.1 Level)

Panel 3: Histogram, Statistically Significant Peaks, Residuals, Version 3, AR≠0 (0.1 Level)

Figure 2: Average Periodogram for Stochastic Variables and Gain Functions for Version 3 and Version 4



PART 8

EDUCATION

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

STUDENT FLOWS AT ONTARIO UNIVERSITIES

Warren Clark

Abstract

Statistics Canada collects data on all university students enrolled as of December of the academic year. Ninety-nine percent of enrolment is collected on an individual student record basis in the University Student Information System (USIS). Student flows showing the progress of students from entry to graduation can be estimated by linking consecutive annual enrolment and degree files. The linking of USIS files also quantifies migration of students between institutions, transitions from one discipline to another and dropout rates. A time series of Ontario university student flows is presented.

1. Introduction

The University Student Information System (USIS) is a data base that provides a Canada-wide system of university enrolment and graduation statistics. Files are available back to the 1972-73 academic year for enrolment and back to 1970 for degrees. The data is collected from university registrars across the country. The data elements include items of educational and academic interest and a wide range of vital characteristics of individual students, such as gender, age, immigration status, country of citizenship, educational activity of the student last year, level, duration of program they are in, number of credits registered in, area of specialization, etc.. All students registered and active for courses which are eligible for academic credit in a degree, diploma or certificate program of a degree-granting institution are included. (Included are students not seeking an academic degree, diploma or certificate if they are taking courses eligible for credit). Ninety-nine percent of enrolment data and 65% of degree data are collected on an individual student record basis.

The USIS system provides a snap-shot picture of enrolment as of December 1st, November 1st for Ontario. The data reported for Ontario universities on USIS is largely reported on an individual student basis for both enrolment and degrees thus enabling the linkage of enrolment and degree files. Only Royal Military College, Ontario Bible College, Ontario Theological Seminary and College Dominicain are excluded from the data reported in this paper because they reported enrolment in aggregate form for some or all of the years examined. These institutions represented 0.7% of Ontario university enrolment in 1988-89.

Most USIS data is collected on an individual student basis with Student ID numbers unique within an institution and in many cases SIN numbers which are unique across the country. It is possible to link student records from one year to the next using these unique identifiers. Thus it is possible to compute the flow of students from one period to a subsequent period. Also because many Ontario universities report their degrees and diplomas granted on an individual student basis it is possible to link the linked enrolment files with the degree files. Those that don't report individual degree records, e.g., Queen's (for some years), Carleton (for some years), Trent and Laurentian had their degree links with the enrolment file imputed based on the number of degrees granted by level and field of study and upon graduation rates for other years or other institutions. Thus not only are student flow percentages calculated but also graduation rates.

The linking of USIS administrative data files permits the comparison of characteristics of students from one year to the next thereby allowing the calculation of flows of students. How many change disciplines? How many move from full-time to part-time studies? How many transfer to other institutions? How many disappear

[1] Warren Clark, Education, Culture and Tourism Division, 16th Floor, R.H. Coats Bldg., Statistics Canada, Ottawa, Ontario K1A 0T6.

from the USIS file (i.e., drop out)? How many switch to a general program from honours or fail to advance to the next year level? These questions can all be answered from the linked files.

Five selected years of linked Ontario university data, covering the academic years from 1976-77 to 1986-87 are presented in this paper: 1976-77, 1979-80, 1981-82, 1983-84, and 1986-87. Each of these USIS enrolment files were linked to the USIS file of the following academic year and to the degree file for that year. (eg. the 1986-87 enrolment file was linked to the 1987-88 enrolment file and to the 1987 degree file).

2. Dropouts

2.1. Introduction

Those who appear on the linked file in the first year but not the second and who do not appear as graduates are called apparent dropouts. The term "apparent" is used to describe dropouts because the linking process is imperfect. Some students who continue their studies are not linked because they changed institutions and could not be traced to their new location. They are misidentified as dropouts. In many cases Social Insurance Numbers are reported on USIS, enabling the tracking of students between institutions. In Ontario in 1986-87 80% of full-time students and 85% of part-time students had a SIN number reported. Those students who do not have a SIN reported and transfer between institutions will be misidentified as dropouts.

The apparent dropout rate represents the percentage of students who drop out between the USIS enrolment count dates of November 1st and November 1st the following year. It does not represent the failure rate of a cohort of students during their university careers. A cohort of new entrants to university may drop out at any time during the 3, 4 or more years they may spend at university. The apparent dropout rate reflects the percentage of dropouts who leave universities without graduating at some time during a year.

Chart 1 Ontario undergraduate enrolment, 1971-1988

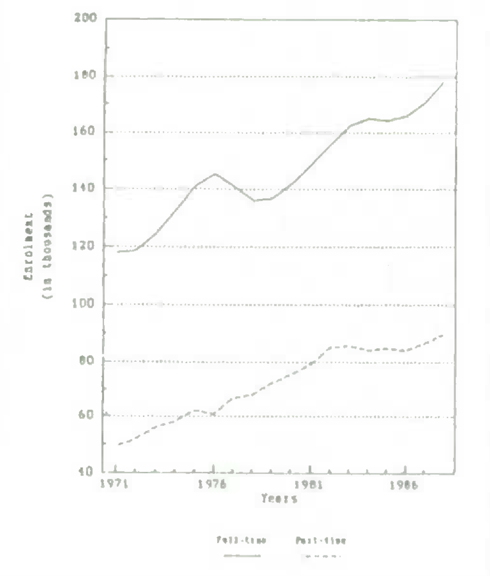
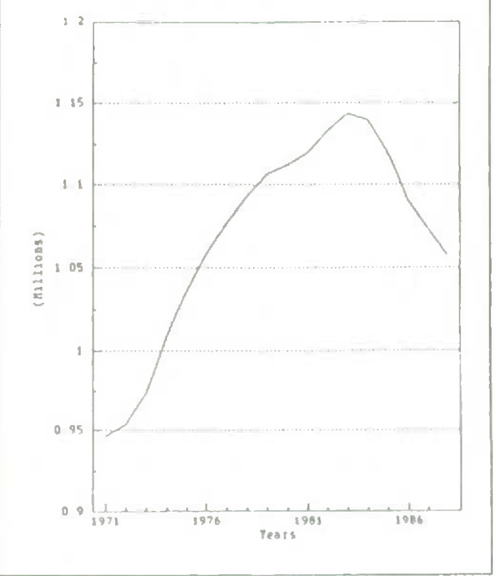


Chart 2 Ontario population age 18 to 24 1971-1988



Ontario undergraduate enrolment has grown as shown in Chart 1. The growth since the early 1980's has occurred despite a decline in the size of the traditional source population for university enrolment, the 18-24 population age group (Chart 2). The increase in enrolment has been the result of:

- a larger percentage of the population completing high school. The number of Ontario Grade 13 graduates related to the 18-year-old population has increase from 28.8% in 1970-71 to 34.5% in 1986-87.

- an increase in the percentage of grade 13 graduates continuing directly on to full-time studies at university. The percentage increased from 65.6% in 1976-77 to 73.1% in 1987-88.

- higher participation of people over age 24 in university education. Between 1971-72 and 1988-89 full-time university enrolment over age 24 increased by 75% whereas 18-24-year-old enrolment increased by 45%. Although enrolment over age 24 increased substantially, it still only represented 19.6% of full-time university enrolment in 1988-89.

- a decline in the dropout rate at Ontario universities

Between 1976-77 and 1986-87 the apparent dropout rate for full-time undergraduates declined from 15.5% to 11.7%. The numbers represent the percentage of full-time undergraduates who do not appear on the enrolment file the following year and who did not graduate. Full-time graduate apparent dropout rates also declined from 20.4% in 1976-77 to 13.9% in 1986-87.

2.2. Dropouts by level

Most undergraduates are in a program that leads to a bachelor's degree normally after 3 or 4 years of study. Less than 4,000 of the nearly 180,000 1986-87 full-time undergraduates are taking courses not leading towards a degree or diploma. They are called "other undergraduates". This group had the highest apparent dropout rate of all full-time undergraduates while those in first professional programs (i.e., M.D., D.D.S., D.V.M, L.L.B.) had the lowest apparent dropout rates.

The part-time undergraduate apparent dropout rate remained relatively stable at about 45% to 48% between 1976-77 and 1986-87. The part-time graduate apparent dropout rate declined from 31% in 1976-77 to 24% in 1986-87. Many of the part-time students identified as apparent dropouts may eventually return to complete their studies.

Table 1. Apparent dropout rate[1] at Ontario universities by registration status and level

Registration status and level	1976-77	1979-80	1981-82	1983-84	1986-87
			(percent)		
Full-time					
Undergraduate	15.5	13.4	12.4	12.4	11.7
Bachelor's	15.1	13.1	11.9	11.8	11.3
First Professional (M.D., D.D.S., D.V.M., L.L.B.)	4.4	3.3	3.0	2.8	2.9
Undergraduate Diploma	22.5	18.1	16.6	17.9	13.1
Other Undergraduates	46.7	45.0	43.9	42.4	39.4
Graduate	20.4	19.3	17.0	14.0	13.9
Master's	16.7	15.9	13.1	9.7	10.0
Doctorate	12.9	11.8	10.6	8.3	8.3
Graduate Diploma	26.5	23.7	20.1	15.5	10.7
Other Graduates	23.0	36.5	30.4	25.4	29.6
Part-time					
Undergraduate	45.5	47.8	47.7	48.3	47.1
Bachelor's	37.4	35.1	35.9	36.7	36.1
Undergraduate Diploma	42.2	53.1	52.4	51.0	54.3
Other Undergraduates	67.9	66.1	64.3	65.4	64.5
Graduate	31.2	30.9	27.1	26.6	24.0
Master's	27.2	28.0	22.8	22.7	20.0
Doctorate	27.5	28.7	25.1	22.8	21.8
Graduate Diploma	47.2	36.8	42.0	36.9	39.0
Other Graduates	55.1	56.8	57.9	57.0	61.0

[1] Apparent dropout rates refer to the percentage of students who appear on the USIS file one year but not the next and who did not graduate. It represents the percentage who apparently drop out or left the Canadian universities without graduating. Students who could not be linked are identified as apparent dropouts.

2.3. Dropouts by field of study

The apparent dropout rates varied considerably depending upon the field of study students were enrolled in as shown in Table 2. General arts and science students experienced the highest apparent dropout rates over the entire time period followed by fine and applied arts students.

Table 2. Apparent dropout rates of full-time undergraduates by major field of study.

Major field of study	1976-77	1979-80	1981-82	1983-84	1986-87
			(percent)		
Total	15.5	13.4	12.4	12.4	11.7
Agriculture and biological sciences	13.0	12.8	13.5	13.6	10.5
Business, management and commerce	16.7	10.6	9.7	9.1	9.4
Education	11.4	9.9	8.9	9.3	7.8
Engineering and applied sciences	15.0	11.7	10.3	9.2	8.7
Fine and applied arts	18.1	17.5	16.0	16.7	14.6
General arts and science	20.4	18.1	16.9	16.9	18.7
Health professions	4.5	4.4	3.2	3.2	3.1
Humanities	16.3	15.6	14.6	13.8	11.5
Mathematics and physical sciences	11.9	11.5	9.8	10.5	10.5
Social sciences	12.8	12.4	12.2	12.4	11.5

Several individual disciplines for full-time undergraduates had much higher than average apparent dropout rates in 1986-87. They were library science, 26.5%; agriculture, 23.3%; public health, 20.3%; and performing arts (excluding music), 19.7%. Library science and the performing arts have had high dropout rates since 1976-77 while agriculture and public health had average or below average rates in 1976-77.

2.4. Dropouts by year level

Table 3 shows that first year students have consistently had the highest apparent dropout rates over the last ten years. Apparent dropout rates at each year level have all declined since 1976-77.

Table 3. Apparent dropout rates of full-time students in 3 or 4 year bachelor's degree programs, by year level

Year level	1976-77	1978-80	1981-82	1983-84	1986-87
			(percent)		
1st year	21.6	18.1	16.9	16.3	15.6
2nd year	15.1	12.8	11.1	11.4	11.4
3rd year	10.4	9.4	8.8	8.6	8.4
4th year	9.0	8.2	7.5	8.3	7.1

In 1976-77, 21.6% of the 1st year students in 1976-77 who apparently dropped out and were not enrolled in 1977-78 and did not graduate. Perhaps more interesting would be the percentage of a cohort of new students who dropout without obtaining a degree. This percentage could be obtained by linking several years of enrolment and degree files together to track a cohort of new students through their careers at university. However, this would be a time-consuming and expensive exercise and rates would not be available until 4 or 5 years after the students had entered, enough time to allow them to graduate. Alternatively, cohort dropouts rates could be estimated by applying current apparent dropout and graduation rates to a fictitious cohort of new students in the following way:

Let E_t^i = the number of students in year t entering year level i

d_t^i = apparent dropout rate in year t and level i (i.e. rates from table 3)

g_t^i = graduation rate from year level i in year t (i.e. the percentage)

D_t = estimated % of a cohort of new entrants to university who dropout before graduation

For all year levels, i,

$$E_t^{i+1} = E_t^i\,(1-d_t^i-g_t^i)$$

$$D_t = \Sigma_i\, d_t^i E_t^i / E_t^1 * 100\%$$

Using this technique, and the apparent dropout rates shown in table 3, the apparent percentage of a cohort dropping out of university without receiving a degree is calculated. Table 4 shows that 43.9% of the 1976-77 new students would dropout compared with 34.3% of the 1986-87 cohort.

Table 4. Cohort dropout rate[1] for full-time 3 or 4 year bachelor's degree students.

1976-77	43.9%
1979-80	41.1%
1981-82	37.4%
1983-84	35.8%
1986-87	34.3%

[1] Percentage of a cohort of new entrants to university who fail to graduate.

2.5. Dropouts who return to university

So far in this paper, the term dropout has referred to students who were not enrolled the following year and who did not graduate, apparent dropouts. With rapidly changing technologies, more and more people are looking upon education as a life long avocation to help them keep pace with rapidly expanding knowledge base. A student may leave and re-enter the education system many times during his/her lifetime. Although anyone who leaves university without graduating is identified as an apparent dropout, they may return to university at some later time. To quantify how many students identified as apparent dropouts actually did return, the 1983-84, 1984-85, 1985-86 enrolment files and 1984 degree file were linked together. Students enrolled in 1983-84, but not in 1984-85 and who did not graduate in 1984 (i.e., apparent dropouts) were matched to the 1985-86 enrolment file. Of the full-time undergraduate apparent dropouts, 21% had returned to their studies in 1985-86. Dropouts from first year were less likely to return than second or third year students.

Table 5. 1983-84 full-time students who apparently dropped out and returned to university in 1985-86 by level and year level

Level and year level	Number of dropouts student	% returning as full-time student	% returning as part-time	% returning to university
Undergraduates				
Total	19,977	14.5	6.5	21.0
1st year	8,756	14.0	4.6	18.6
2nd year	5,295	19.5	7.1	26.7
3rd year	3,168	12.8	10.4	23.2
4th year	1,515	10.4	8.4	18.8
5th year	24	12.5	8.3	20.8
Not applicable	1,182	6.3	4.9	11.3
Graduates				
Total	2,889	7.2	4.6	11.8
Master's	1,081	6.9	9.0	15.9
Doctorate	472	9.3	4.9	14.2
Graduate Diploma	46	8.7	2.2	10.9
Other Graduate	96	6.3	6.3	12.5

3. Student Flows

3.1. Introduction

The linking process not only identifies those students who are not found on the file the following year, dropouts, but it enables the comparison of student characteristics from one year to the next for those who are found. For example, students may change levels of study, disciplines, year levels, institutions or change from full-time to part-time study or vice versa. The linked USIS files in essence permits the tracking of students or the measurement of the flow of students from one educational state to another.

3.2. Migration from level to level

With declining apparent dropout rates for full-time undergraduates, slightly higher percentages of them remain as full-time undergraduates or transfered to part-time undergraduate studies in 1986-87 than occurred in 1976-77. The 1986-87 full-time undergraduates did not show any greater tendency to move on to full-time graduate studies than the 1976-77 cohort, however, they were less likely to continue to part-time graduate studies. The 1986-87 part-time undergraduate cohort were less likely to remain in part-time studies and more likely to change to full-time than the 1976-77 cohort. Full-time graduates were more likely to remain as full-time graduates and less likely to change to part-time studies in 1986-87 than in 1976-77.

Table 6. Flows of students between levels at Ontario universities

Registration status and level	Total	Activity the following year					
		Not enrolled		Enrolled[1]			
		Apparent Dropouts	Graduated[2]	Undergraduate		Graduate	
				Full-time	Part-time	Full-time	Part-time
		Percent of Total					
Full-time undergraduate							
1976-77	142,576	15.5	16.0	61.2	5.7	1.6	0.2
1979-80	135,463	13.4	14.9	64.0	6.1	1.6	0.2
1981-82	147,365	12.4	13.8	65.1	6.7	1.7	0.2
1983-84	161,106	12.4	14.3	58.9	6.9	1.2	0.1
1986-87	164,465	11.7	15.1	64.5	7.0	1.6	0.1
Part-time undergraduate							
1976-77	59,960	45.5	7.2	5.0	41.3	0.5	0.6
1979-80	71,615	47.8	7.4	6.3	38.6	0.6	0.6
1981-82	78,639	47.7	6.5	5.6	38.8	0.6	0.6
1983-84	85,323	48.3	7.1	7.2	36.1	0.6	0.6
1986-87	83,726	47.1	7.9	7.4	36.4	0.6	0.6
Full-time graduate							
1976-77	18,298	20.4	16.7	1.6	0.5	51.4	9.4
1979-80	17,817	19.3	15.0	1.2	0.6	53.9	10.0
1981-82	19,196	17.0	16.5	1.1	0.8	55.6	9.1
1983-84	20,636	14.0	17.4	0.9	0.7	58.2	8.8
1986-87	21,381	13.9	17.7	1.1	0.6	58.3	8.4
Part-time graduate							
1976-77	12,316	31.2	14.7	0.5	1.2	3.7	48.7
1979-80	12,041	30.9	14.3	0.4	2.3	4.3	47.9
1981-82	11,839	27.1	16.7	0.5	4.5	2.3	48.9
1983-84	11,862	26.6	16.9	0.5	1.9	4.6	49.4
1986-87	11,455	24.0	19.2	0.4	1.6	5.0	49.8

[1] Includes students who graduated and continued their studies.
[2] Includes those who graduated but did not re-enrol.

Note: Percentages may not add to 100% due to rounding.

3.3. Transitions from one discipline to another

During their time at university students may change majors several times. Regardless of the registration status or level of study the vast majority of students remained in the same discipline from one year to the next. Table 7 shows the discipline transfer rates by major field of study (i.e., it indicates what percentage of students transfer to another discipline from one year to the next, eg., from psychology to sociology, etc.). Table 7 indicates that between 1986-87 and 1987-88 23.0% of full-time undergraduates changed the discipline they were studying. This compared with 19.5% in 1976-77 and 27.1% in 1983-84. Many students may start out in general arts and science and eventually decided to study a particular discipline. For this reason general arts and science students are the most likely to change disciplines.

The near doubling of the general arts and science percentage between 1981-82 and 1983-84 is the result of a change in coding methods of disciplines at a large university. Before 1983-84 nearly all undergraduates at that institution were coded as general arts and/or science students regardless of the discipline they were studying. Only some highly specialized disciplines were coded separately. Thus most undergraduates were identified as staying in general arts or science for all their undergraduate years. In 1983-84 that institution began to code students who were formerly coded as arts or science into individual disciplines. Those that remained in the general arts and science category could at any time during their academic career transfer to an individual discipline which would also be reflected in the coding of majors. Now that these transfers were being recorded, the transfer rate nearly doubled from 33.8% in 1981-82 to 62.1% in 1983-84. This an artificial increase created by the change in coding methods used. Prior to 1983-84 the transfer rates for general arts and science would have likely been close to what they are today had the current coding system been in place

Table 7. Discipline transfer rates[1] by major field of study, level and registration status.

	1976-77	1979-80	1981-82	1983-84	1986-87
			(Percent)		
Full-time undergraduate					
Total	19.5	20.3	20.9	27.1	23.0
Agriculture and biological sciences	17.3	18.2	20.0	17.6	17.8
Business, management and commerce	10.8	14.6	18.5	19.2	15.5
Education	10.1	10.6	13.1	11.6	11.6
Engineering and applied sciences	16.9	17.2	17.5	14.2	16.0
Fine and applied arts	9.7	11.5	12.8	10.6	11.4
General arts and science	30.7	35.6	33.8	62.1	60.8
Health professions	7.1	7.9	8.4	8.8	9.6
Humanities	17.0	15.4	16.3	15.9	15.3
Mathematics and physical sciences	16.5	15.2	15.3	16.3	15.1
Social sciences	13.7	14.0	15.3	13.4	14.0
Full-time graduate					
Total	4.9	7.8	5.2	4.6	4.9
Agriculture and biological sciences	5.1	11.5	3.7	4.9	3.4
Business, management and commerce	6.0	4.5	3.5	3.1	2.1
Education	9.7	4.1	5.8	3.6	3.8
Engineering and applied sciences	1.7	3.6	2.3	1.6	1.8
Fine and applied arts	3.0	2.0	3.5	2.2	1.4
General arts and science	3.7	..	10.3	4.3	5.0
Health professions	7.6	12.3	8.9	6.1	7.6
Humanities	5.7	4.6	4.3	5.5	3.9
Mathematics and physical sciences	2.2	7.9	1.4	2.1	2.1
Social sciences	2.8	2.6	5.6	5.4	2.7

[1] Percentage of students who change disciplines from one year to the next (e.g., from psychology to sociology, from mathematics to physics, etc.)

3.4. Sources of students

So far the focus of this paper has been on where do students go from one year to the next. Do they change disciplines, levels or do they apparently drop out. It is possible to examine this data in another way; where do students come from? Are they new students (i.e., not enrolled last year), were they part-time students last year, were they enrolled at the same level or a lower level? The linked USIS files can answer all these questions. For example of the 1987-88 Ontario full-time master's students, 21% were undergraduates the previous year, 44% were master's students, 2% were at another level of graduate studies, and 32% were not enrolled in 1986-87. Of full-time doctoral students in 1987-88, 0.5% were undergraduates, 14% were master's students, 71% were doctoral students and 0.7% were at other levels of graduate studies in 1986-87.

3.5. Migration between institutions

When SIN numbers are reported inter-institutional transfers can be tracked. In Ontario where most transfers occur between institutions within the province and where 80% of full-time and 85% of part-time enrolment have SIN numbers reported, the vast majority of inter-institutional transfers can be tracked. Tracking transfers outside Ontario is more difficult, particularly to institutions in Quebec, because SIN number reporting is not as common. Those that cannot be tracked due to non-reporting of SIN numbers result in students being misidentified as apparent dropouts even though they continued their studies at another institution. Table 8 shows that less than 4% of full-time undergraduates and 2.4% of full-time graduates changed institutions between 1986-87 and 1987-88. The percentages have not varied much over the ten years from 1976-77 to 1986-87.

Table 8. Inter-institutional transfers (Percentage of students who transferred between universities from one year to the next)

Registration status and level	1976-77		1979-80		1981-82		1983-84		1986-87	
	to an institution		to an institution		to an institution		to an institution		to an institution	
	Within Ontario	Outside Ontario	Within Ontario	Outside Ontario	Within Ontario	Outside Ontario	Within Ontario	Outside Ontario	Within Ontario	Outside Ontario
					(Percent)					
Full-time:										
Undergraduate	4.1	0.3	3.5	0.5	4.1	0.5	3.4	0.4	3.4	0.4
Graduate	2.0	0.3	1.6	0.5	2.2	0.7	1.8	0.5	1.9	0.5
Part-time:										
Undergraduate	2.8	0.1	3.4	0.3	4.2	0.4	3.7	0.2	3.7	0.3
Graduate	1.2	0.1	1.7	0.3	2.0	0.3	1.9	0.2	1.5	0.3

Inter-institutional transfers occurred more frequently when students changed registration status (full-time/part-time) or when they changed levels of education (Table 9). For example, in 1986-87 30% of

Table 9. Inter-institutional transfers by change in registration status and level from one year to the next

Registration status and level Year t	Registration status and level Year t+1	Year t 1976-77	1979-80	1981-82	1983-84	1986-87
				(percent)		
Full-time undergraduate	→Full-time undergraduate	4	4	4	4	3
Full-time undergraduate	→Full-time graduate	31	29	29	30	30
Full-time undergraduate	→Part-time undergraduate	12	12	13	11	10
Full-time undergraduate	→Part-time graduate	30	28	37	30	32
Full-time graduate	→Full-time graduate	2	2	3	2	2
Full-time graduate	→Part-time graduate	2	2	1	1	2
Full-time master's	→Full-time doctorate	13	13	10	11	11
Part-time undergraduate	→Full-time undergraduate	14	18	17	12	12
Part-time undergraduate	→Full-time graduate	31	37	39	42	39
Part-time undergraduate	→Part-time undergraduate	4	6	7	7	6
Part-time undergraduate	→Part-time graduate	31	30	29	30	34
Part-time graduate	→Full-time graduate	7	10	10	9	9
Part-time graduate	→Part-time graduate	1	1	1	1	1

students who advanced from full-time undergraduate to full-time graduate studies changed institutions compared with 3% who stayed as full-time undergraduates and 10% who changed to part-time undergraduate studies. Six percent of part-time undergraduates who stayed as part-time undergraduates changed institutions but 12% of those who switched to full-time studies changed.

4. Limitations

4.1. Students leaving the country

Students who leave the country to study abroad are misidentified as dropouts. In 1986 approximately 17,000 Canadians were studying abroad at postsecondary institutions (universities and community colleges). The number who came from Canadian universities the previous year, and therefore were misidentified as dropouts, is unknown.

4.2. Transfers to community college

Students who transfer to community colleges or institutes of technology before completing their university education are also identified as dropouts. The Ontario College Information System indicates that approximately 1,500 students came from universities the previous year. This means that about 1% of full-time undergraduates are classified as dropouts even though they continued their studies at community college.

4.3. Transfers to other universities in Canada

Inter-institutional transfers can only observed where SIN numbers are reported (e.g. transfers from University of Toronto which reports SIN numbers to McGill which does not report SIN numbers could not be tracked and students would be identified as dropouts). Approximately 4% of students change institutions from one year to the next. Ontario has a high SIN response rate. This enables the tracking of students between institutions within Ontario. However, because Quebec has a relatively low SIN response rate transfers from Ontario institutions to Quebec institutions are unlikely to be traced and will be misidentified as apparent dropouts. Examination of the previous educational activity of Ontario residents studying outside Ontario in 1986-87 indicates that at most 2,300 students who were classified as apparent dropouts may actually have been transfers to institutions outside the province which could not be traced. This represents 1.3% of students misidentified as apparent dropouts who were actually transfers to institutions outside the province. Since 80% of full-time undergraduates in Ontario report SIN and 4% of students were identified as changing institutions, if 100% reported SIN, then 5% of students may have changed institution, thereby reducing the apparent dropout rate by another 1%.

4.4. Fall semester only

USIS looks at enrolment at only one point in time during the academic year, December or November. Approximately 5% to 10% of all students do not attend during the fall semester but do during the winter, spring or summer semesters. Thus this group of students is missed by USIS.

4.5. Dropouts before the USIS count date

Students enrolling for the first time in September but who dropout before the USIS enrolment count in December or November are not captured by USIS and are not identified as dropouts.

5. Applications to other areas of Canada

The same methodology could be applied to British Columbia, Manitoba and New Brunswick where the incidence of SIN number reporting is high and where most degrees are reported on a an individual student basis (Table 10). The other provinces either have the majority of their degrees reported in aggregate form thereby precluding the identification of individual graduating students, or have a low percentage of SIN reporting thereby not permitting the tracking of inter-institutional transfers.

Agreements between provincial governments in the Maritime provinces which allow one university in the Maritimes to offer specialized programs for all students in the region has encouraged transfers between institutions. Inter-institutional transfers are much more common than in Ontario. Unfortunately, due to the low incidence of SIN reporting in Nova Scotia and Prince Edward Island it is impossible to accurately track student flows between institutions in the Maritimes. This inability to track inter-institutional transfer would result in many continuing New Brunswick students who transfer to Nova Scotia or Prince Edward Island to be misidentified as dropouts.

Table 10. Applications to other provinces

Province	% of undergraduate degrees and diplomas reported on an individual record basis	% of full-time students reporting S.I.N.
Newfoundland	--	94
Prince Edward Island	82	--
Nova Scotia	56	33
New Brunswick	91	83
Quebec	7	27
Ontario	87	80
Manitoba	88	72
Saskatchewan	30	30
Alberta	55	53
British Columbia	70	92

Note: Provinces with high percentages in both columns are good candidates to use linked files to estimate dropout rates and student flows.

6. Conclusion

The linking of USIS files provides valuable information on the flows of students, dropout rates and the sources of students, for educational researchers, manpower planners and university administrators. In Ontario the apparent dropout rate for university students has fallen over the ten year time period from 1976-77 to 1986-87 which has been a contributing factor to continued growth of undergraduate enrolment despite declines in the 18-24 population age group. Apparent dropout rates and student flows can be calculated based on any characteristic of students contained on the USIS enrolment file.

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

THE FEMALE/MALE EARNINGS GAP AMONG RECENT UNIVERSITY GRADUATES: THE FIRST FIVE YEARS

Ted Wannell[1]

ABSTRACT

This report focuses on the female/male earnings differential of a very select group -- 1982 university graduates who were employed full-time in 1984 or 1987. Although the earnings gap in this young, well-educated cohort was smaller than that found in the general working population, female graduates earned less than males in almost every category examined. Furthermore, the gap widened from 1984 to 1987. A multivariate model is introduced to better control for the many factors affecting the earnings gap. Despite the controls, the model accounted for just one-third of the earnings gap at each timepoint.

INTRODUCTION

The existence of an earnings gap between men and women is not news. Until recently, traditional family roles dictated a division of labour within the household such that the majority of married women were engaged in unpaid household and child-rearing activities. Since most women's paid working career ended at marriage or the birth of their first child, most women did not accumulate the skills and experience necessary to climb the salary ladder. Thus a wide gulf existed between the salaries of working men and women.

But times change. A combination of social, demographic and economic trends have resulted in the increased participation of women in the full-time workforce. Fewer women are leaving the labour force when they marry or have children. Interruptions for childbirth have also been reduced by the long-term drop in fertility and the costs are subsidized through the Unemployment Insurance plan. As more and more women entered and remained in full-time paid jobs, the salary gap narrowed, yet it remains substantial.

In 1987, females working full-time for the full year earned, on average, a third less than their male counterparts. Of course, a gap of this size reflects many differences between the female and male workforce: age structure, education, occupation, industry of employment and accumulated experience. The waters are furthered muddied by the changes in these variables over time. Today's labour force consists of many cohorts, each with a unique set of characteristics, which entered the labour market under very different conditions.

What if we could follow a single recent cohort of labour market entrants about whom most of the important income-related characteristics were known? Would we find that men and women with the same qualifications were earning about the same? These are precisely the types of issues that can be addressed with the National Graduates Survey and the Follow-up of Graduates Survey.
The National Graduates Survey of 1984 (NGS) and the Follow-up of Graduates Survey in 1987 (FOG) yield a unique perspective on the recent status of the female / male earnings gap. The sampling frame for these surveys encompasses the 1982 graduates of all universities in Canada.[1] The surveys captured a wide range of demographic, educational and labour market information covering the period from 1982 to 1987. Making use of these surveys, the intent of this paper is to look at two issues:

[1] Ted Wannell, Business and Labour Market Analysis Group, Analytical Studies Branch, Statistics Canada, Ottawa, Ontario. The author wishes to acknowledge Monica Boyd, Marie-Claire Couture, Joanne Dubeau, Doug Giddings, Bill Magnus, Yigal Messeri, Debbie O'Dwyer, Garnett Picot and Wendy Pyper.

i. given recent increases in female labour force participation and educational attainment, does an earnings gap still exist for men and women with equal qualifications; and,

ii. how does the earnings gap develop over time within a particular cohort.

While some descriptive statistics and cross-tabulations are presented, most of the analysis is of a multivariate nature. A technique known as decomposition is used divide the earnings gap into a component that can be explained by differences in education and background characteristics and a residual component that could be indicative of some relative disadvantages for one of the groups under study. Decomposition results are presented for earnings in 1984 and 1987.

The NGS/FOG data indicate that a substantial gap exists between the earnings of recent male and female university graduates. Although narrowed within some categories, the gap is persistent across almost all fields and levels of study. Only about one-third of the gap could be explained by differences in the educational and background characteristics of men and women. The earnings gap was also found to grow over time for the cohort under study.

A Brief Note on the Data

The National Graduates Survey (NGS), 1984, and the Follow-up of Graduates Survey (FOG), 1987, collected a wide range of information on the labour market experiences of 1982 community college and university graduates. Included on each survey was a question asking respondents to estimate their yearly earnings (to the nearest thousand dollars) based on the job held at the time of the interview. Responses to this question typically contain two digits (ie. 10-99 thousand). To avoid any representation of spurious accuracy most figures in this paper are also reported in two digit numbers.

The analysis in this report is limited to full-time workers. The descriptive comparisons of 1984 earnings include all graduates working full-time in 1984 and a similar restriction applies to the comparison of 1987 earnings. Since the number of hours worked is highly variable for part-time workers and the surveys did not query the number of hours worked, this condition ensures that approximately equal amounts of labour are being compared. It also follows that the earnings figures approximate full-time, full-year earnings due to the way the question was asked. The full-time restriction yields the following maximum sample sizes for the descriptive tables:

	1984	1987
Male	5141	4986
Female	4032	3689
Total	9173	8675

The exact sample sizes for each table will be somewhat smaller due to missing values for the variables under study.

A much more restrictive definition was used in the multivariate analyses: only those employed full-time at each of 5 separate timepoints were included. The resultant maximum sample is 5971 (3582 males and 2389 females). Working samples are substantially smaller due to missing values among the many variables included in the analysis. Where possible, the descriptive comparisons were produced for the regression population and vice versa. Neither set of results was substantially altered by changing the population specifications.

More detailed information on the surveys is available from the Household Surveys Division in the form of users' guides and methodology reports.

Background

The last 30 years have witnessed dramatic change in the Canadian labour market -- most notably the increasing participation of women in full-time jobs. Between 1967 and 1988, the proportion of full-time jobs held by women rose from 27 percent to almost 39 percent. This proportion should continue to move upwards as the full-time participation rates of younger women are higher than in older cohorts.

Gains in the educational attainment of women are even more dramatic than increasing labour force participation. While only a quarter of undergraduate university degrees went to women in the early 1960s, women made up more than half the graduates in the late 1980s. Women still lag behind at the graduate level, but are catching up rapidly. The female share of masters degrees rose from 19 percent in 1961 to 45 percent in 1989. Women accounted for less than a tenth of earned doctorates in 1961, but nearly a third by 1989.

Despite the increasing participation in full-time work and gains in educational attainment, the earnings of women still lag well behind those of men. While the earnings gap has narrowed in the past 20 years it remains substantial. In 1987, women working full-time for the full-year earned, on average, a third less than their male counterparts. Can differences in the age structure, education or experience of the male and female workers account for such a large gap? To answer that question for the entire labour force would be very difficult indeed, but the NGS/FOG panel opens a window on a recent cohort of labour market entrants about whom most of the relevant characteristics are known. In this section, we will examine the earnings gap among increasingly specific groups of university graduates from the panel.

Looking at the 1982 university graduates, females employed full-time in 1984 earned an average of 24 thousand -- or 87 percent of the male average of 27 thousand. By 1987, the ratio of female to male earnings had dropped to 82 percent, with female earnings averaging 31 thousand compared to 38 thousand for males.

The earnings gap was smaller among the university graduates than the full-time workforce of approximately the same age. Among the general working population the same age as the 1982 university graduates, females earnings averaged 18 thousand and males 25 thousand -- a ratio of 70 percent.[2] Similarly in 1987, the age-weighted female workforce comparable to the university graduates earned 71 percent of the male average. Age-weighting tends to narrow the earnings gap because male and female wages are closer together in the younger age groups which contain the majority of graduates.

The divergent field of study distributions of males and females are striking. Many fields tend to be dominated by one sex or the other. The differing field of study distribution can affect the overall earnings gap. If men tend to gravitate to high-reward fields of study, the gap could be inflated. A simple method to check for this bias is to compare the within-field earnings ratios to the overall ratio. If the average of the within-field ratios is significantly lower than the overall ratio, differing field of study choices by men and women account for some proportion of the earnings gap.

The within-field female / male earnings ratio averaged 89 percent in 1984 and 85 percent in 1987, narrowing the overall earnings gap by 2 and 3 percentage points, respectively.

Even though the earnings gap is generally smaller within fields of study, women graduates of virtually all programs still earn less than men. In fact, in only one field -- Political Science -- did female graduates earn at least as much as men in 1984. But the earnings pendulum swang back in favour of the men by 1987.

Degree level is another variable that accounts for some salary stratification. University graduates with doctorates employed full-time in 1984 earned 45 percent more (35 percent more in 1987) than those with undergraduate degrees, with masters-level graduates occupying the middle ground. The earnings gap was largest among masters graduates, with ratios of 85 percent in 1984 and 81 percent in 1987, compared to 90 percent and 83 percent for undergraduates.

Table 1. Female to Male Earnings Ratios 1982 University Graduates Employed Full-time in 1984 or 1987 by Field of Study

Field of Study	Female/Male Ratio 1984	Female/Male Ratio 1987
	%	%
1. Education	87	86
2. Fine Arts	96	89*
3. Applied Arts	..	..
4. Journalism	..	..
5. Other Humanities	98	94
6. Sociology, Anthropology, Demography	99	97
7. Criminology	..	..
8. Law	88	95
9. Economics	88	75
10. Geography/Environment	83	82
11. Political Science	104	86
12. Psychology	83	82
13. Other Social Sciences	90	86*
14. Agriculture	..	..
15. Biochemistry, Biology, Zoology	90	95
16. Home Economics	..	..
17. Veterinary	..	..
18. Architecture	..	..
19. Engineering	89	89*
20. Forestry	..	..
21. Lanscape Architecture	..	..
22. Dentistry	..	..
23. Medicine	81	87
24. Nursing	..	..
25. Optometry	..	..
26. Pharmacy	..	..
27. Public Health	..	..
28. Computer Sciences	95	91*
29. Math	97	93*
30. Chemistry, Geology, Metallurgy		84
31. Meteorology	90	..
32. Physics/Other	..	..
Unweighted Average	89	85
Weighted Average	87	82

Note: .. - sample size too small to publish (coefficient of variation > 25%)

* - relatively small sample size, interpret cautiously (coefficient of variation: 16.5% - 25%)

Table 2. Female to Male Earnings Ratios 1982 University Graduates Employed Full-Time in 1984 or 1987, by Field of Study and Program Length

Field of Study	Program Length	Female/Male Ratio 84	Female/Male Ratio 87
		%	%
All Fields	Undergrad Univ	90	83
	Master/Grad Cer	85	81
	Doctorate	101	99
Education	Undergrad Univ	92	89
	Master/Grad Cert	83	86
	Doctorate	91	88
Fine Arts and Humanities	Undergrad Univ	99	91
	Master/Grad Cert	95	95
	Doctorate	105	94
Commerce, Economics and Law	Undergrad Univ	87	87
	Master/Grad Cert	87	89
	Doctorate	..	..
Other Social Sciences	Undergrad Univ	94	90
	Master/Grad Cert	89	84
	Doctorate	93	91
Agriculture and Biological Sciences	Undergrad Univ	91	80
	Master/Grad Cert	89	84
	Doctorate	87	89
Engineering	Undergrad Univ	91*	89*
	Master/Grad Cert	80*	..
	Doctorate	111	119
Medical and Health Sciences	Undergrad Univ	65	54
	Master/Grad Cert	77	50
	Doctorate	158*	118*
Math and Physical Sciences	Undergrad Univ	95	93
	Master/Grad Cert	83*	89*
	Doctorate	94*	94*
Unweighted Average		94	92
Weighted Average		87	82

Note: .. - sample size too small to publish (coefficient of variation > 25%)

* - relatively small sample size, interpret cautiously (coefficient of variation: 16.5% - 25%)

The gap is virtually nonexistent at the doctorate level: women Ph.D.s earned 1 percent more than men in 1984 and 1 percent less in 1987.

Combining the effects of field of study and degree level should then narrow the earnings gap. And this is indeed the case. Averaged across 10 major fields of study and three degree levels, female university graduates earned 94 percent of the salaries of their male counterparts in 1984 and 92 percent in 1987. Of course, this average is disproportionately influenced by Ph.D. holders who make up only a small proportion of the population.

Furthermore, field of study aggregations at this level run into some problems. For example, the Medical and Health Sciences category compares a male population consisting mostly of M.D.s to a female population of mainly nursing graduates. The detail (or homogeneity of groups) in cross-tabulations is limited by the high variability of small within-cell sample sizes.

Since descriptive analysis is limited to one or two variables, it is conceivable that the combined effects of other variables may account for some within-cell differences in the earnings of men and women. Accordingly, the next section examines the simultaneous effects of many variables on the earnings gap using a multivariate technique.

Decomposing the Female / Male Earnings Differential

In the previous section, the earnings gap between male and female graduates was categorized by only one or two variables simultaneously. While further cross-classifications or more detailed categories might create more comparable groups, small within-cell sample sizes severely limit the range of such analyses. On the other hand, a multivariate approach allows the effects of a number of variables to be studied simultaneously and the results to be assessed by standard test scores. In this section, a multivariate technique known as decomposition is used to analyze the gender differential in earnings.

The decomposition technique is based upon linear regressions of the earnings of two different groups; in this case, male and female graduates. The regression equations are structured on a human capital model: earnings are modelled as a function of education and experience (investment in human capital), while controlling for background or demographic characteristics. As was noted in the previous section, at least some of the earnings gap is due to differences in the courses of study chosen by men and women. The same may hold true for experience or any of a range of background characteristics. The decomposition technique is primarily a tool for estimating the proportion of the earnings gap attributable to the measured differences in human capital and background characteristics of men and women. The remaining difference in earnings is referred to as the residual component. The regression coefficients allow the residual difference to be subdivided into differential returns individuals' characteristics.

It is important to remember that the decomposition results are estimates subject to both specification and measurement error. The results can be affected by unmeasured human capital characteristics or self-selection (e.g. a **graduate's** choice of one occupation over another for non-monetary reasons). Accordingly, decomposition cannot provide direct evidence of wage discrimination. On the other hand, it can suggest which characteristics might be differentially rewarded. A brief description of the decomposition methodology follows.

Methodology

The 'non-discriminatory' decomposition technique outlined by Cotton (1988) is employed in this paper. This technique is a variant of a methodology that dates back to the 1950s and has appeared in economic, sociological and demographic literature.[3]

Consider the following earnings equation:

$$\ln W = b X + u$$

where $\ln W$ is the natural log of yearly earnings[4]; X is a (k,j) matrix of k observed data values for j variables, b is a vector of j coefficients measuring returns to those variables and u is the error term. Identical earnings equations are estimated for sub-samples of men and women using ordinary least squares (OLS). Once the coefficients are estimated, the error term drops out, the OLS estimators b replace b and the superscripts m and f identify the male and female equations; resulting in

$$\ln W^f = \hat{b}^f X^f$$

$$\ln W^m = \hat{b}^m X^m.$$

One property of OLS estimators is that the product of the coefficients and associated variable means sum to the mean of the independent (left-side) variable, so that

$$\overline{\ln W^f} = \hat{b}^f \overline{X^f}$$

$$\overline{\ln W^m} = \hat{b}^m \overline{X^m}.$$

The decomposition technique centres on the fact that the difference in mean earnings is a simple function of explanatory variable means and the **estimated** return to these characteristics. Therefore, if men and women received the same return to their endowments b^f equals b^m and the difference in earnings would be solely attributable to differing endowments.

Cotton recognized that in the absence of differential treatment, the return to endowments would fall somewhere between those for the currently advantaged and disadvantaged. He proposes that these 'non-discriminatory' coefficients be estimated as the weighted average of the male and female coefficients. Therefore,

$$\hat{b}^* = p^m \hat{b}^m + p^f \hat{b}^f \ ,$$

where $\hat{b}^*$ is the vector of non-discriminatory coefficients and p^m and p^f are the proportions of the total population that are female and male.

With several simple steps that need not be repeated here, Cotton arrives at a decomposition of the earnings differential containing three terms:

$$\begin{aligned} \overline{\ln W^m} - \overline{\ln W^f} &= \hat{b}^* (\overline{X}^m - \overline{X}^f) \\ &+ \overline{X}^m (\hat{b}^m - \hat{b}^*) \\ &+ \overline{X}^f (\hat{b}^* - \hat{b}^f). \end{aligned}$$

The first term represents the component of the earnings gap attributable to differing endowments (human capital and background characteristics). The second and third terms divide the residual earnings differential into male treatment advantage (higher than expected return to endowments) and female treatment disadvantage (lower than expected return to endowments).

The dollar values expressed for these components sum to the difference in the **geometric** mean earnings of men and women, as opposed to the arithmetic mean used in the previous section. The geometric mean is simply the anti-log of the average log earnings. That is

$${}_{e}(\overline{\ln W}).$$

The equations were estimated for earnings 1984 and 1987, and the change in earnings between 1984 and 1987. **The population was limited to graduates with valid earnings data in 1984 and 1987 who were working full-time at each of the five time points covered in the surveys.**[5] Thus the subpopulations of females and males have a history of strong and more-or-less equal attachment to the labour force. This definition yields a conservative estimate of the earnings gap as more women have interrupted work histories or work part-time.[6] The sensitivity of the results to alternative population definitions was tested and will be discussed later.

The independent (explanatory) variables include controls for age, language, province, interprovincial mobility, parents' postsecondary education, marital status, children, work experience prior to studies, detailed field of study, degree level and public sector employment. For the 1984 and 1987 equations, most variables are specific to each timepoint. For the earnings change equation, the 1984 controls are entered as well as any change that takes place in those controls between 1984 and 1987.

Note that industry and occupation controls are not included in the model. The process of matching graduates to jobs in different industries and occupations may be conditioned on sex. Therefore, controls for industry and occupation may mask one element of earnings discrimination (Gunderson, 1988). On the other hand, a public sector employment control is included since the wide implementation of target group programs and the stated merit principle of hiring and advancement in that sector may create a separate set of job matching rules.

Results

The differing human capital and background characteristics (endowments) of men and women accounted for relatively little of the earnings gap in 1984 and 1987. This 'explained' proportion accounts for about a third of the earnings gap at each timepoint. Differing field of study patterns for each sex accounted for most of the 'explained' proportion of the gap. Public sector employment was also an important factor, but usually acted in the opposite direction to field of study -- it tended to be an equalizing characteristic. The residual difference was usually a fairly even split between the male treatment advantage and female treatment disadvantage components.[7]

The geometric mean earnings of men exceeded the female mean by $3700 in 1984 and $7000 in 1987 (see Table 3). In both years, the differing endowments of men and women accounted for 35 percent of the earnings gap (i.e. $1300 in 1984 and $2500 in 1987). Divergent field of study distributions were the most important factors -- making up 133 percent of the net difference in 1984 and 84 percent in 1987.[8] The higher percentage of men with masters degrees was also a large factor in 1984, but less so in 1987. Age (in both years) and public sector employment (in 1984) were the strongest factors narrowing the explained earnings gap. The women in the population were, on average, older than the men, with age yielding positive returns. Similarly, a higher percentage of females worked in the public sector in both years. However, public sector employment yielded positive returns in 1984 but negative returns in 1987 -- largely due to a high premium accruing to men for working in the private sector in 1987. Therefore, the higher percentage of women in the public sector was an equalizing variable in 1984, but helped to widen the earnings gap in 1987. In a similar vein, women earned higher returns to previous full-time work experience (particularly 3 years and over) in 1984, but this effect dissipated by 1987.

Table 3. Estimated Components of the Female/Male Earnings Gap Among 1982 University Graduates[1] 1984 and 1987[2]

	1984		1987	
	$	%	$	%
Differing Characteristics	1,300	35	2,500	36
Differing Returns to Characteristics				
- Male Advantage	1,000	28	2,300	33
- Female Disadvantage	1,400	37	2,200	31
Total Earnings Gap[3]	3,700	100	7,000	100

[1] Employed full-time at all 5 timepoints: January 1983, October 1983, June 1984, January 1986 and March 1987.

[2] Detailed decomposition results are available from the author.

[3] Difference in the geometric average earnings of males and females.

Discussion

The NGS/FOG data indicate that a sizable earnings gap exists between recent male and female university graduates and that the gap grows over time. Furthermore, the gap persists across almost all fields and levels of study, Ph.D.s being the exception to the rule. Through the use of a multivariate model, it was estimated that only about a third of the earnings gap in 1984 and 1987 could be explained by differences in the education or background of men and women.

What about the remaining, unaccounted for differences in the earnings of male and female graduates? Is this evidence of discrimination? Or are there alternative explanations? Obviously, there are no cut and dried answers. Let's consider three general classes of explanations for the residual difference: omission of explanatory variables, incomplete measurement of the dependent variable (income), and demand side discrimination.

Omitted Explanatory Variables

The regression models that fed into the decomposition calculations accounted for between 30 and 50 percent of the total variation in earnings of the graduates. While such fit statistics compare favourably with most microdata-based earnings models, the fit might have been improved by unmeasured or improperly measured earnings-related variables. However, to contribute to the explained proportion of the earnings gap, one group would need to have appreciably more (or less) of this unobserved characteristic. That condition makes some possibilities hard to fathom.

Some combination of latent ability or ambition is often cited as a possible source of unexplained variance in earnings models. It doesn't seem a very reasonable proposition that such characteristics should be unequally distributed between large subgroups of the population (i.e. men possessing more than women), particularly with very specific education controls already in the model. If anything, one would expect that the female graduates are more highly selected for latent ability, since a smaller proportion met the conditions to be included in the sample.[9] In other words, the women who consistently worked full-time over the period under study were generally those with better income-generating characteristics than those with interrupted or part-time work patterns. The distinction was not so clear among the male graduates.

One point that was clear from an earlier analysis was that male graduates tended to enter higher-paying occupations than female graduates, even when controlling for field of study (Wannell, 1989). Occupation (and/or industry) could have been included among the control variables but, as was noted earlier, the differential matching of male and female graduates to first jobs could involve some discriminatory processes.[10] Such barriers to the hiring or advancement of women would probably play as much of a role in a comprehensive model of discrimination as pay differentials within narrow (but not necessarily homogenous) occupational categories.

On the other hand, there may be some tendency for women to choose jobs differently than men. There are many reasons for this: the teaching of sex roles in the education and socialization process; a preference to work or deal with other women; or even the perception (whether justified or not) that women generally fare better in some occupations or sectors than others. This may at least partially explain the much higher percentage of women in the public service where there is less of an earnings gap. Ironically though, the slow earnings growth in the public sector exacerbates the earnings gap over time. In any case, such an effect would be difficult to isolate in a single equation model of earnings. A simultaneous equations approach, combining models of occupation (or industry) selection and earnings, would be more appropriate. In such a milieu, one would also have to consider the possibility of the incomplete measurement of income.

Incomplete Measurement of the Dependent Variable (Income)

Earnings, as measured by the NGS and FOG, is just one component (though usually the largest) of the total income from any job. Total income, in its broadest sense, also includes both monetary and non-monetary benefits. Monetary benefits include such things as company contributions to pension plans, dental plans and supplemental medical plans. Non-monetary benefits are less easily defined or measured: job satisfaction, geographic preferences, challenge or the opportunities for advancement.

Unfortunately, the NGS/FOG contains no information on the benefits received by respondents. While it is unlikely that the distribution of benefits would be so balanced in the favour of women as to significantly narrow the earnings gap, the different types of jobs held by men and women could allow for some small differences. If, for example, benefits were greater in the public sector than the private sector, the gap -- as measured by earnings alone -- might be slightly overestimated.

The issue of non-monetary rewards is closely linked to the discussion of occupational choice in the previous section. The earnings gap, in this line of argument, may be narrowed if women receive greater non-monetary returns -- such as satisfaction -- from their jobs. The NGS/FOG offer some evidence in this area, but it is not clear how it should be interpreted. Both surveys asked respondents to rate their overall satisfaction with their jobs and, more specifically, with their salaries. Men and women were equally satisfied with their jobs, indicating that women are probably not receiving greater non-monetary rewards than men. The salary satisfaction question did, however, show that slightly more women than men (22% compared to 19%) were dissatisfied with their salaries.

Demand Side Discrimination

Sex discrimination by employers can be classified into two basic forms: hiring and advancement selection based on sex; and, differential pay for the same work. Each form of discrimination presents its own problems to researchers. The existence of hiring and advancement selection has stronger theoretical underpinnings and seems to fit NGS/FOG data. However, employer selection is so functionally and theoretically similar to self-selection (choice) that it is virtually impossible to distinguish the two. On the other hand, unequal pay for the same work does not have strong theoretical support, is probably identifiable only with a case study format and may be difficult to isolate from selection effects.

At the risk of oversimplification, most theoretical discussions of hiring and advancement selection boil down to 'statistical discrimination'.[11] To briefly summarize this argument, women are more likely to interrupt their working careers for marriage and child care than men. Employers prefer 'career-track' employees who do not have short or frequently interrupted careers. Since employers cannot readily determine at the time of hiring which women will have short or interrupted careers, hiring or advancing a man is a better bet, particularly in jobs that require significant on-the-job training and career development, all other considerations being equal. While this type of discrimination has traditionally been illustrated by the shunting of women to 'pink collar ghettos' such as clerical work, it is not necessarily that blunt. There is enough leeway within most highly-qualified occupations for some stratification by sex to occur. But at this level the distinction between discrimination and self-selection is not always clear.

Human capital theory suggests that women who plan to have shorter or interrupted careers would favour jobs that have relatively shorter periods of on-the-job training and offer little penalty for time spent out of the labour force. Trade-offs may exist within most occupations whereby some jobs can be exited and re-entered relatively easily, but with some pay or benefit penalties. The self-selection of different occupational streams reinforces the earnings gap, but is not normally labelled discrimination.

Of course the distinction between discrimination and self-selection is blurred by other factors such as childhood instruction in male and female roles, the expectation of discrimination in some occupations and the individual's imperfect knowledge of their own future. All of which makes it very difficult to clearly identify hiring and advancement discrimination.

The issue of earnings differentials for essentially the same duties is even more difficult to isolate. Regardless of the problems with theoretical arguments, this phenomenon simply cannot be measured with normal survey microdata. There are two main reasons for this. The first involves sampling ratios. Surveys typically sample only a small proportion of the population, therefore making it highly unlikely that men and women with similar qualifications doing similar jobs at the same firm could be identified. Even though the NGS/FOG provide many controls for qualification and experience and have a very low sampling ratio, the jobs held by the graduates represent a minute fraction of all jobs in the labour market. Even if a few matches could be found (from which no statistical inferences could be drawn), the second problem would come into play: occupation coding.

Canadian microdata sources -- NGS/FOG included -- at best contain occupation information coded at the Standard Occupation Classification four-digit level. What this jargon means is that the entire range of jobs in Canada is summarized into less than 500 categories. Obviously jobs within categories cannot be entirely homogenous at this level. If homogenous jobs cannot be identified, then unequal pay for the same job cannot be measured. Furthermore, job titles may be the source of discrimination. It is possible that essentially similar job duties may be given different titles or descriptions for men and women (Baron and Bielby, 1986). Focused case studies may provide some insights, but bring about a different set of problems.

One final point to consider is that the wage gap among graduates is small compared to other groups in the labour force and disappears altogether for Ph.D.'s. If male and female earnings are converging over time (that is, for subsequent cohorts), isolating and analyzing a wage gap will become increasingly difficult. Joint labour supply decisions and fertility may deserve more attention than a diminishing earnings gap.

References

Belman, Dale and John S. Heywood, Government wage differentials: a sample selection approach, Applied Economics, **21**, 1989, 427-438.

Bielby, William and James Baron, Men and women at work: sex segregation and statistical discrimination, American Journal of Sociology, **91**, 1986, 759-799.

Cotton, Jeremiah, On the decomposition of wage differentials, Review of Economics and Statistics, **70**-2, 1988, 236-243.

England, Paula, The failure of human capital theory to explain occupational sex segregation, Journal of Human Resources, **17**, 1982, 358-370.

Finney, Ross, The Gender Wage Gap and Job Experience in a Simultaneous Equations Framework: A Look at Younger Workers, presented at the 23rd Annual Meeting of the Canadian Economics Association, June, 1989.

Fishback, Price V. and Joseph V. Terza, Are estimates of sex discrimination by employers robust? The use of never-marrieds, Economic Inquiry, **28**-2, 1989, 271-285.

Glass, Jennifer, Marta Tienda and Shelley A. Smith, The impact of changing employment opportunity on gender and ethnic earnings inequality, Social Science Research, **17**, 1988, 252-276.

Gunderson, Morley, Male-female wage differentials and policy es, Journal of Economic Literature, 27, March, 1989, 46-72.

Heckman, James J., The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Annals of Economic and Social Measurement, 5, 1976, 479-492.

Jackson, John D. and James T. Lindley, Measuring the extent of wage discrimination: a statistical test and a caveat, Applied Economics, 21, 1989, 515-540.

Jones, F. L., On decomposing the wage gap: a critical comment on Blinder's method, Journal of Human Resources, 18, 1983, 126-130.

Mincer, Jacob and Solomon Polachek, Family investments in human capital: earnings of Women, Journal of Political Economy, 82, 1974, S76-108.

Mincer, Jacob and Solomon Polachek, Womens's earnings re-examined, Journal of Human Resources, 13, 1978, 113-134.

Reimers, Cordelia, Labour market discrimination against hispanic and black men, Review of Economic and Statistics, 65-3, 1983, 570-579.

Robb, Roberta, Earnings differentials between males and females in Ontario, 1971, Canadian Journal of Economics, 11, 1978, 350-359.

Statistics Canada, Household Surveys Division, Follow-up of Graduates: Survey Methodology Report and User's Guide, March 1989.

Statistics Canada, Education, Culture and Tourism Division, The class of 82 revisited, Education Statistics Bulletin, 11-1, 1989.

Thurow, Lester, Generating Inequality: Mechanisms of Distribution in the United States, New York: Basic, 1975.

Wannell, Ted, The persistent gap: Exploring the earnings differential between recent male and female postsecondary graduates, Employment and Immigration Canada / Statistics Canada, mimeo, August, 1989.

Endnotes

[1] The graduates of community colleges and vocation and trade programs were also surveyed but are not included in this report for two reasons. First, the entrance requirements vary greatly, thus the graduates are much less homogenous group with respect to age, years of education and academic skills. Second, trade and vocational (and, to a lesser extent community college) programs are so stratified by sex that female to male comparisons are subject to high sampling variability (due to small numbers in the minority group). A longer version of this paper includes analyses of the earnings gap among community college graduates and is available, upon request from the author.

[2] Age-weighted average earnings are calculated by multiplying the average earnings for a particular age group (from **Earnings of Men and Women**, Statistics Canada Catalogue 13-217) by the proportion of graduates in the age group and summing across age groups.

3. For a more complete attribution of the history of the technique, see Cotton (1988) or Gunderson (1988).

4. The natural log of earnings is used so that the estimated coefficients approximate the proportionate effect on earnings of changes in the variables on the right-hand side of the equation (Gunderson, 1988).

5. These time points are: January 1983, October 1983, June 1984, January 1986 and March 1987. These criteria open the possibility of sample selection bias. Tests with a Heckman-type correction for this bias indicate that the 'explained' component of the earnings gap is probably overestimated in the absence of the correction. In other words, we are less likely to find evidence of discrimination in an uncorrected decomposition.

6. For example, 33 percent of the female graduates met these criteria to be included in the multivariate analysis, compared to 41 percent of the males.

7. An alternative decomposition, proposed by Jackson and Lindley (1989), facilitates an F-test of the significance of the residual component. According to this procedure, the residual components of all three decompositions -- earnings 1984, earnings 1987 and the change in earnings -- are significant.

8. Since some variables can have the opposite effect to the overall trend (i.e. favour females), the absolute sum of differences can easily exceed the net sum of differences. The subtotal of 133 percent for fields of study in 1984 indicates that other field of study, women had 'better' wage-generating characteristics than men (e.g. higher average age, more public sector employment, etc.).

9. If the sample selection procedure is treated as a selectivity bias problem with a Heckman-type correction, the explained proportion of the 1987 earnings gap drops from 35 percent to 15 percent. In other words, the evidence of discrimination becomes stronger.

10. The 1987 decomposition calculations were repeated for a model containing 15 occupation and 12 industry dummy variables. The explained component of the earnings gap rose from 36% to 47%. Clearly men and women with similar qualifications are getting different types of jobs. To what extent the differences are supply or demand side generated is very difficult to determine.

11. For a longer, more general discussion on statistical discrimination see Thurow (1975).

CLOSING REMARKS

Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

CLOSING REMARKS

David A. Binder[1]

One of the objectives of holding this symposium was to heighten awareness among both theoretical and applied statisticians of the special problems which are raised when data are analyzed over time, as opposed to an analysis of cross-sectional data at a single point in time. By including both theoretical papers as well as particular examples of data analyses which have been performed, we have succeeded in bringing together a group where there has been worthwhile cross-fertilization of knowledge and ideas.

By browsing through the papers presented, we see that analysis of data in time touches on a very diverse set of applications. We have seen examples which have been derived from areas such as:

- manufacturing establishments and other business surveys
- censuses
- income distributions
- labour force status
- gender earnings gap
- demographic rates
- educational attainment and student flows
- cardiovascular disease
- HIV/AIDS
- cancer mortality
- housing prices
- supermarkets
- air pollution
- soil conservation and land use

There are a few common threads which are repeated throughout these papers. I will concentrate here on those aspects where the time dimension poses new challenges which are not present in the analysis of cross-sectional data.

Measurement Error

Measurement errors are present in virtually every statistical and observational process. These arise for a number of reasons, including misunderstanding the concepts, respondent's inability to provide an accurate response, transcription and processing errors, deliberately answering incorrectly, etc. In cross-sectional studies these errors are often assumed to occur at random with little or no impact on the bias of the estimates. However, in analysis of data in time, where an estimate of change is often at least as important as an estimate of level, measurement errors can have a substantial impact. For example, estimates of changes in classification (e.g. labour force status) can often be significantly affected by a relatively small measurement error.

Therefore, when designing a study where measurement of change or behaviour over time is important, more attention should be paid to reducing measurement errors and possibly adjusting the data to account for measurement error.

Adjustment Procedures

Related to measurement errors are other factors which demand special procedures to adjust the data prior to analysis. For example underreporting and under-coverage of the frame, where the level of error decreases over time for a particular reference period, may require methods where the error levels are modelled and the data are adjusted. Other adjustment procedures for comparing

[1] David A. Binder, Assistant Director, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

populations at two or more time points may also be required. These can be due to seasonal effects, business cycle effects, and changes in the population mix due to age composition, ethnic origin mixes, frame changes, etc. If such changes in the mix of a dynamic population can be expected to lead to different characteristics of the population, it is important to account for these in the analysis.

As well, the implications on the survey design are clear. It is important to have sufficient information on these population mixtures, so that such an adjustment can be performed. As in the case of any analysis, the analyst needs to be aware of the various factors which can help explain the population behaviour.

Changes in Concepts

Finally, one of the most difficult challenges in the analysis of data over time is the problem where concepts being measured have been altered from one time point to the next. This can occur for a variety of reasons. The previous concept may now be out of date, e.g. head of household vs. household maintainer, or quality change of a product used in a price index. An administrative data base may have undergone a change in the administrative system. There are many examples of this phenomenon. The options open to the analyst are limited. It is often not feasible or even advisable to maintain the old definitions. This will be a constant concern to data analysts. Of course, the first step to resolving these issues is to recognize that the problem exists. Resolving this problem would normally be handled on a case-by-case basis.

In closing, the conference organizers and editors of these proceedings should be thanked for their efforts in holding such a successful symposium. Although many have contributed to its success, special thanks are due to Patricia Whitridge and Avi Singh for co-ordinating all the activities.

Most importantly, though, special thanks are due to all the presenters and authors who have contributed to this symposium and who have ensured its success.

De même, les conséquences sont claires pour ce qui a trait à la conception de l'enquête. Il est essentiel que nous ayons suffisamment de données sur la composition de la population pour pouvoir exécuter des redressements. Comme dans n'importe quelle analyse, l'observateur doit connaître les divers facteurs qui peuvent expliquer l'évolution démographique.

Modification des concepts

Enfin, l'un des problèmes les plus épineux de l'analyse chronologique est le fait que des définitions peuvent être modifiées dans l'intervalle. Cela peut s'expliquer de diverses façons. La définition peut, par exemple, être devenue désuete (par ex. : soutien du ménage au lieu de chef du ménage, changement de qualité pour un produit qui sert à la construction d'un indice de prix). Ou encore, on peut avoir fait subir à une base de données administratives des changements fondamentaux. Nous pourrions citer beaucoup d'autres exemples de ce phénomène. En revanche, les solutions qui s'offrent à l'analyste sont limitées. Il peut rarement conserver l'ancienne définition, soit parce que cela est impossible ou que ce n'est pas recommandé. Les analystes de données ne doivent jamais perdre cela de vue. Reconnaître l'existence du problème est évidemment la condition préalable pour le résoudre. A cette fin, on procédera normalement par étude de cas.

En terminant, je me dois de remercier les organisateurs du symposium et les éditeurs des actes pour les efforts qu'ils ont investis dans la préparation de ce congrès. Bien que de nombreuses personnes aient contribué au succès de ce symposium, je dois remercier tout particulièrement Patricia Whitridge et Avi Singh pour avoir coordonné toutes les activités.

Je remercie également de façon toute particulière les conférenciers et les auteurs qui ont participé à ce symposium et qui en ont fait un succès.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

DISCOURS DE CLOTURE

David A. Binder[1]

Le symposium auquel nous venons de participer avait notamment pour but de sensibiliser davantage les spécialistes de la statistique théorique et appliquée aux problèmes particuliers qui se posent dans l'analyse de données chronologiques et que l'on ne retrouve pas dans l'analyse de données transversales. En invitant les participants à nous soumettre des communications théoriques ou à nous présenter des exemples d'analyse de données, nous avons pu réunir des personnes qui se sont livrées à un échange fructueux de connaissances et d'idées.

En parcourant les articles qui ont été présentés à ce symposium, nous voyons que l'analyse des données dans le temps recouvre des sujets très variés. Les exemples qui nous ont été présentés portaient sur des sujets comme:

- l'enquête sur les manufactures et les autres enquêtes- entreprises
- le recensement
- la répartition du revenu
- la situation vis-à-vis de l'activité
- la différence de rémunération entre les hommes et les femmes
- les taux démographiques
- le niveau de scolarité et les flux d'étudiants
- les maladies cardio-vasculaires
- le VIH et le SIDA
- la mortalité par cancer
- le prix du logement
- les supermarchés
- la pollution atmosphérique
- la conservation et l'utilisation du sol.

Tous ces articles renferment quelques idées communes. Je ferai plutôt ressortir ici les aspects sous lesquels la dimension temporelle pose des difficultés que l'on ne retrouve pas dans l'analyse de données transversales.

Erreur de mesure

Il y a des erreurs de mesure dans presque toutes les enquêtes statistiques. Ces erreurs surviennent pour diverses raisons : mauvaise compréhension des concepts, incapacité du répondant de donner une réponse juste, erreurs de transcription et de traitement, malhonnêteté du répondant, etc. Dans les analyses transversales, on suppose souvent que ces erreurs surviennent de façon aléatoire et qu'elles n'ont à peu près aucun effet sur le biais des estimations. Toutefois, dans les analyses chronologiques, où une estimation de variation est souvent au moins aussi importante qu'une estimation de niveau, les erreurs de mesure peuvent avoir une incidence appréciable. Par exemple, les estimations des variations de l'effectif d'une catégorie (par ex. : la situation vis-à-vis de l'activité) peuvent souvent être modifiées de façon significative à cause d'une erreur de mesure relativement faible.

Par conséquent, lorsqu'on élabore une enquête où la mesure des variations ou de l'évolution chronologique occupe une place importante, on doit s'attacher davantage à réduire la probabilité d'erreur de mesure et à prévoir un mécanisme de redressement des données pour tenir compte de ce genre d'erreurs.

Méthodes de redressement

D'autres facteurs liés aux erreurs de mesure exigent des opérations spéciales visant à redresser les données avant analyse. Par exemple, dans les cas de sous-déclaration ou de sous- dénombrement, où le niveau d'erreur pour une période de référence donnée diminue avec le temps, on pourrait devoir modéliser les niveaux d'erreur et redresser les données. On pourrait aussi devoir recourir à d'autres méthodes de redressement pour comparer des chiffres de population à des périodes différentes. Cela pourrait être nécessaire à cause des effets saisonniers, des effets des cycles économiques ou de changements dans la composition de la population attribuables à une modification de la structure par âge, du mode de répartition des origines ethniques, des bases de sondage, etc. Si on prévoit que de tels changements peuvent modifier les caractéristiques de la population, il importe d'en tenir compte dans l'analyse.

[1] David A. Binder, directeur adjoint, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario), K1A 0T6.

ALLOCUTION DE CLÔTURE

aux études. Deuxièmement, il y a une telle stratification, selon le sexe, dans les programmes de formation professionnelle ainsi que dans les cours de métier (et, jusqu'à un certain point, dans les collèges communautaires) que les comparaisons entre les femmes et les hommes sont sujettes à une variabilité d'échantillonnage élevée (à cause du petit nombre de personnes dans le groupe minoritaire). Une version plus longue du présent article comprend des analyses de l'écart entre les gains parmi les diplômés des collèges communautaires. On peut se procurer ce dernier document en s'adressant à l'auteur.

[2] Pour obtenir les gains moyens pondérés en fonction de l'âge, on multiplie les gains moyens pour un groupe d'âges particulier (tirés de **Gains des hommes et des femmes**, publication n^{o} 13-217 au catalogue de Statistique Canada) par la proportion de diplômés dans ce groupe d'âges puis on fait la somme de ces résultats pour tous les groupes d'âges.

[3] Pour plus de détails sur l'historique de cette technique, voir Cotton (1988) ou Gunderson (1988).

[4] Nous utilisons le logarithme naturel des gains afin que les coefficients estimés donnent une approximation de l'effet proportionnel qu'ont, sur les gains, les changements dans les variables de la partie droite de l'équation (Gunderson, 1988).

[5] Ces périodes de référence sont: janvier 1983, octobre 1983, juin 1984, janvier 1986 et mars 1987. Ces critères pourraient donner lieu à un biais dans la sélection de l'échantillon. Des tests, comprenant une correction de type Heckman, effectués en rapport avec ce biais montrent que la composante 'expliquée' de l'écart entre les gains est probablement **surestimée** quand on n'applique pas la correction. En d'autres mots, il est moins probable que nous trouvions des preuves de discrimination dans une décomposition non corrigée.

[6] Par exemple, 33 pour cent des femmes détenant un diplôme répondaient aux critères pour être incluses dans l'analyse multidimensionnelle, comparativement à 41 pour cent des hommes.

[7] Une autre décomposition, proposée par Jackson et Lindley (1989), facilite la réalisation d'un test F afin de déterminer si la composante résiduelle est significative. Selon cette procédure, les composantes résiduelles des trois décompositions -- gains en 1984, gains en 1987 et changement dans les gains -- sont significatives.

[8] Puisque certaines variables peuvent avoir un effet contraire à celui de la tendance globale (c.-à-d. qu'elles peuvent favoriser les femmes), la somme des valeurs absolues des différences peut facilement dépasser la somme nette des différences. Le total partiel de 133 pour cent pour les domaines d'études en 1984 montre que, mis à part le domaine d'études, les femmes avaient de 'meilleures' caractéristiques qui influent sur le traitement que ce n'était le cas pour les hommes (par ex., un âge moyen plus élevé, un taux d'emploi plus élevé dans le secteur public, etc.)

[9] Si l'on traite la procédure de sélection de l'échantillon comme un problème de biais de sélectivité avec une correction de type Heckman, la proportion expliquée de l'écart entre les gains en 1987 tombe de 35 pour cent à 15 pour cent. En d'autres mots, le preuve de discrimination devient plus évidente.

[10] Les calculs de décomposition pour 1987 ont été repris pour un modèle contenant 15 variables accessoires relatives à la profession et 12 relatives à la branche d'activité. La composante expliquée de l'écart entre les gains a augmenté de 36% à 47%. Il est manifeste que les hommes et les femmes avec des compétences semblables obtiennent des genres d'emplois différents. Il est très difficile de déterminer dans quelle mesure les différences constatées sont produites du côté de l'offre ou du côté de la demande.

[11] Pour une étude plus longue et plus générale sur la discrimination statistique, voir Thurow (1975).

fonctions du poste sont essentiellement semblables (Baron et Bielby, 1986). Des études de cas centrées pourront donner une certaine compréhension intuitive, mais elles amèneront un autre ensemble de problèmes.

Un dernier point à considérer est le fait que l'écart entre les salaires, parmi les diplômés, est petit comparativement à celui qui existe parmi d'autres groupes dans la population active et il disparaît complètement pour les détenteurs d'un doctorat. Si les gains des hommes et ceux des femmes convergent dans le temps (c'est-à-dire pour les cohortes suivantes), il deviendra de plus en plus difficile d'isoler et d'analyser un écart entre les gains. Il se peut que les décisions conjointes relatives à l'offre de travail et à la fécondité méritent plus d'attention qu'un écart entre les gains qui diminue.

BIBLIOGRAPHIE

Belman, Dale and John S. Heywood, Government wage differentials: a sample selection approach, *Applied Economics*, 21, 1989, 427-438.

Bielby, William and James Baron, Men and women at work: sex segregation and statistical discrimination, American *Journal of Sociology*, *91*, 1986, 759-799.

Cotton, Jeremiah, On the decomposition of wage differentials, *Review of Economics and Statistics*, *70-2*, 1988, 236-243.

England, Paula, The failure of human capital theory to explain occupational sex segregation, *Journal of Human Resources*, *17* 1982, 358-370.

Finney, Ross, The Gender Wage Gap and Job Experience in a Simultaneous Equations Framework: A Look at Younger Workers, presented at the 23rd Annual Meeting of the Canadian Economics Association, June, 1989.

Fishback, Price V. and Joseph V. Terza, Are estimates of sex discrimination by employers robust? The use of never-marrieds, *Economic Inquiry*, *28-2*, 1989, 271-285.

Glass, Jennifer, Marta Tienda and Shelley A. Smith, The impact of changing employment opportunity on gender and ethnic earnings inequality, *Social Science Research*, *17*, 1988, 252-276.

Gunderson, Morley, Male-female wage differentials and policies , *Journal of Economic Literature*, 27, March, 1989, 46-72.

Heckman, James J., The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement*, 5, 1976, 479-492.

Jackstone, John D. and James T. Lindley, Measuring the extent of wage discrimination: a statistical test and a caveat, *Applied Economics*, *21*, 1989, 515-540.

Jones, F.L., On decomposing the wage gap: a critical comment on Blinder's method, *Journal of Human Resources*, *13*, 1978, 113-134.

Reimers, Cordelia, Labour market discrimination against hispanic and black men, *Review of Economic and Statistics*, *65-3*, 1983, 570-579.

Robb, Roberta, Earnings differentials between males and females in Ontario, 1971, *Canadian Journal of Economics*, *11*, 1978, 3050-359.

Statistics Canada, Household Surveys Division, Follow-up of Graduates: Survey Methodology Report and User's Guide, March 1989.

Thurow, Lester, *Generating Inequality: Mechanisms of Distribution in the United States*, New York: Basic,, 1975.

Wannell, Ted, The persistent gap: Exploring the earnings differential between recent male and female postsecondary graduates, Employment and Immigration Canada / Statistics Canada, mimeo, August, 1989.

NOTES EXPLICATIVES

[1] Les enquêtes ont aussi recueilli des données sur les diplômés des programmes de formation professionnelle et des cours de métier, mais ces personnes ne sont pas incluses dans le présent rapport pour deux raisons. Premièrement, les conditions d'admission varient beaucoup, les diplômés de ces cours forment donc un groupe beaucoup moins homogène pour ce qui est de l'âge, du nombre d'années de scolarité et des aptitudes

réduit si les femmes reçoivent, de leur emploi, des récompenses non monétaires supérieures -- comme la satisfaction. L'END et l'ESD confirment jusqu'à un certain point cette hypothèse, mais l'interprétation de cette constatation n'est pas claire. Dans le cadre des deux enquêtes, on a demandé aux enquêtés d'évaluer comment ils sont satisfaits globalement de leur emploi et, plus particulièrement, de leur salaire. Les hommes et les femmes étaient également satisfaits de leur emploi, ce qui montre que les femmes ne reçoivent probablement pas plus de récompenses non monétaires que ce n'est le cas pour les hommes. Cependant,la question portant sur la satisfaction en matière de salaire a permis de trouver qu'un peu plus de femmes que d'hommes (22% comparativement à 19%) étaient mécontentes de leur salaire.

Discrimination du point de vue de la demande

La discrimination, faite par les employeurs, en fonction du sexe peut être classée selon deux formes de base: la sélection, en matière d'embauchage ainsi que d'avancement, basée sur le sexe et une rémunération différente pour le même travail. Chaque forme de discrimination présente ses propres problèmes aux chercheurs. L'existence de la sélection en matière d'embauchage ainsi que d'avancement est mieux étayée du point de vue théorique et elle semble s'accorder avec les données de l'END et de l'ESD. Cependant, la sélection faite par les employeurs ressemble tellement, de façon fonctionnelle et théorique, au libre choix qu'il est virtuellement impossible de faire la différence entre les deux. Par contre, il n'existe pas d'argument théorique sérieux qui appuie le fait qu'une rémunération différente serait versée pour le même travail, ce n'est probablement que sous forme d'études de cas que l'on pourrait reconnaître l'existence de cette discrimination et il pourrait être difficile de l'isoler des effets reliés à la sélection.

Au risque de faire une simplification excessive, on peut dire que la majorité des études théoriques portant sur la sélection en matière d'embauchage et d'avancement se ramènent à une 'discrimination statistique'. Résumons brièvement cet argument, il est plus probable que les femmes plutôt que les hommes interrompent leur carrière suite à leur mariage ou pour garder leurs enfants. Les employeurs préfèrent les employés qui ont un cheminement de carrière et dont la carrière n'est pas brève ou interrompue fréquemment. Puisque les employeurs ne peuvent reconnaître facilement, au moment de l'embauchage, les femmes dont la carrière sera brève ou interrompue fréquemment, il leur est plus profitable d'embaucher un homme ou de donner de l'avancement à un homme, particulièrement dans le cas d'emplois pour lesquels il faut beaucoup de formation pratique et de promotion de la carrière, toutes autres choses étant égales par ailleurs. Bien que ce genre de discrimination ait traditionnellement été illustré par le fait que les femmes étaient aiguillées vers des 'ghettos de cols roses', comme le travail de bureau, elle n'est pas nécessairement aussi brutale. Il y a assez de marge de manoeuvre dans la majorité des professions de haute qualification pour qu'il se produise une certaine stratification selon le sexe. Mais à ce niveau, la distinction entre la discrimination et le libre choix n'est pas toujours évidente.

Selon la théorie du capital humain, les femmes qui prévoient avoir une carrière brève ou interrompue favoriseraient les emplois où les périodes de formation en cours d'emploi sont plus courtes et où il y a peu de pénalité pour les périodes passées à l'extérieur de la population active. Il se peut qu'il existe, dans la majorité des professions, des mécanismes de compensation qui permettent de quitter certains emplois et de les occuper à nouveau assez facilement, mais avec des pénalités relatives au salaire ou aux avantages sociaux. Le libre choix de différentes professions renforce l'écart entre les gains, mais, dans ce cas, on ne parle pas, normalement, de discrimination.

Bien entendu, la distinction entre la discrimination et le libre choix est brouillée par d'autres facteurs tels que la formation reçue dans l'enfance quant aux rôles masculins et féminins, la discrimination anticipée dans certaines professions ainsi que la connaissance imparfaite qu'a chaque personne de son propre avenir. Tous ces éléments font qu'il est très difficile de reconnaître clairement la discrimination en matière d'embauchage et d'avancement.

Il est encore plus difficile de faire ressortir les différences entre les gains pour des fonctions essentiellement identiques. Quels que soient les problèmes relatifs aux arguments théoriques, ce phénomène ne peut tout simplement pas être mesuré à l'aide des microdonnées recueillies lors d'une enquête courante. Et ce, pour deux raisons principales: La première porte sur les fractions de sondage. D'ordinaire, les enquêtes ne sont effectuées qu'auprès d'une petite partie de la population, ce qui rend donc très peu probable que l'on puisse identifier des hommes et des femmes avec les mêmes compétences qui remplissent des emplois semblables dans la même entreprise. Bien que l'END et l'ESD comprennent de nombreux moyens pour tenir compte des compétences ainsi que de l'expérience et que le dénominateur de leur fraction de sondage soit très faible, les emplois occupés par les diplômés ne représentent qu'une partie infime de tous les emplois sur le marché du travail. Même si l'on pouvait trouver quelques jumelages (à partir desquels on ne pourrait faire aucune déduction statistique), le second problème, celui du codage des professions, entrerait en jeu.

Les sources canadiennes de microdonnées -- l'END et l'ESD incluses -- renferment, au mieux, des renseignements sur les professions codés selon la Classification type des professions au niveau à quatre chiffres. Ce jargon signifie que la gamme complète de tous les emplois au Canada est résumée en moins de 500 catégories. Il est évident que les emplois dans une catégorie ne peuvent être entièrement homogènes à ce niveau. Si l'on ne peut identifier des emplois homogènes, il ne nous est pas possible de mesurer un salaire inégal pour le même emploi. De plus, les titres d'emplois peuvent être la source de discrimination. On peut donner des descriptions ou des titres différents à des emplois, remplis par des hommes et par des femmes, alors que les d'un doctorat constituant l'exception à la règle. L'emploi d'un modèle multidimensionnel a permis d'estimer que seulement environ le tiers de l'écart entre les gains en 1984 et en 1987 pouvait être expliqué par des différences dans l'instruction ou dans les antécédents des hommes et des femmes.

Qu'en est-il des différences inexpliquées qui restent entre les gains des diplômés et des diplômées? Est-ce que cela constitue une preuve de discrimination? Ou y a-t-il une autre explication? Il n'y a évidemment pas de réponses toutes faites. Considérons trois classes générales d'explications pour la différence résiduelle: l'omission de variables explicatives, la mesure incomplète de la variable dépendante (revenu) et la discrimination du point de vue de la demande.

Variables explicatives omises

Les modèles de régression dont les résultats ont servi aux calculs de décomposition ont expliqué entre 30 et 50 pour cent de la variation totale dans les gains des diplômés des deux sexes. Bien que de telles données d'ajustement se comparent favorablement avec la majorité des modèles des gains basés sur des microdonnées, il est possible que l'on aurait pu améliorer l'ajustement si l'on avait utilisé des variables non mesurées ou mal mesurées relatives aux gains. Cependant, pour contribuer à la proportion expliquée de l'écart entre les gains, il faudrait qu'un groupe possède sensiblement plus (ou moins) de cette caractéristique qui n'a pas été observée. Cette condition rend certaines possibilités difficiles à comprendre.

On cite souvent une certaine combinaison d'aptitudes ou d'ambition latente comme une source possible de différence inexpliquée entre les modèles des gains. Il ne semble pas qu'il soit très raisonnable de supposer que de telles caractéristiques devraient être réparties de façon inégale entre de grands sous-groupes de la population (c.-à-d. les hommes possédant plus que les femmes) particulièrement compte tenu des variables de contrôle très précises en matière d'instruction qui font déjà partie du modèle. On s'attendrait plutôt à ce que les femmes qui détiennent un diplôme universitaire aient fait l'objet d'une sélection plus stricte pour ce qui est des aptitudes latentes, car une plus petite proportion d'entre elles ont satisfait aux conditions pour être incluses dans l'échantillon.[9] En d'autres mots, les femmes qui ont toujours travaillé à plein temps pendant la période visée par l'étude étaient généralement celles qui possédaient de meilleures caractéristiques qui influencent sur les traitements que celles dont la carrière a été interrompue ou qui ont travaillé à temps partiel. La distinction n'était pas aussi claire dans le cas des diplômés de sexe masculin.

Une analyse effectuée auparavant a permis de déterminer que les hommes détenant un diplôme tendaient plus à entrer dans des professions mieux rémunérées que ce n'est le cas pour les diplômées, même quand on tient compte du domaine d'études (Wannell, 1989). La profession et (ou) la branche d'activité aurait pu être incluse dans les variables de contrôle mais, comme on l'a déjà fait remarquer, le jumelage différent des hommes et des femmes détenant un diplôme universitaire avec le premier emploi pourrait comporter certains processus discriminatoires.[10] De telles barrières à l'embauchage ou à l'avancement des femmes joueraient probablement un rôle aussi important dans un modèle de discrimination complet que les différences dans la rémunération à l'intérieur de catégories professionnelles limitées (mais pas nécessairement homogènes).

Par contre, il se peut que la façon dont les femmes choisissent des emplois diffère de celle des hommes. Cela est attribuable à de nombreuses raisons: l'enseignement des rôles sexuels dans le processus d'éducation et de socialisation; une préférence pour travailler ou pour traiter avec d'autres femmes; ou même la perception (qu'elle soit justifiée ou non) qu'en général les femmes réussissent mieux dans certaines professions ou dans certains secteurs que dans d'autres. Cela peut expliquer, du moins partiellement, le pourcentage beaucoup plus élevé de femmes dans la fonction publique où l'écart entre les gains est plus faible. Il est cependant ironique de constater que le faible taux de croissance des gains dans le secteur public exacerbe l'écart entre les gains dans le temps. De toute manière, il serait difficile d'isoler un tel effet à l'aide d'un modèle des gains à une seule équation. Une méthode utilisant des équations simultanées, combinant des modèles de sélection et de gains relatifs à la profession (ou à la branche d'activité), serait plus appropriée. Dans un tel milieu, il faudrait aussi tenir compte de la possibilité d'une mesure incomplète du revenu.

Mesure incomplète de la variable dépendante (revenu)

La variable gains, telle que mesurée par l'END et par l'ESD, n'est qu'une des composantes (bien que ce soit habituellement la plus importante) du revenu total provenant de tout emploi. Le revenu total, pris dans son sens le plus large, inclut aussi les avantages monétaires et les avantages non monétaires. Les premiers comprennent, entre autres, les contributions de l'employeur à un régime de retraite, à un régime d'assurance-soins dentaires et à un régime médical complémentaire. Les avantages non monétaires sont plus difficiles à mesurer ou à définir: la satisfaction professionnelle, les préférences géographiques, le défi ou les possibilités d'avancement.

Malheureusement, les données de l'END et l'ESD ne renferment aucun renseignement sur les avantages reçus par les répondants. Bien qu'il soit peu probable que la répartition des avantages favorise suffisamment les femmes pour réduire de façon significative l'écart entre les gains, les différents genres d'emplois détenus par les hommes et par les femmes pourraient permettre de petites différences. Si, par exemple, les avantages étaient plus importants dans le secteur public que dans le secteur privé, l'écart -- tel que mesuré par les gains seulement -- pourrait être surestimé légèrement.

La question des récompenses non monétaires est étroitement liée à la discussion sur le choix d'une profession qui a été faite dans la section précédente. Selon ce raisonnement, il se peut que l'écart entre les gains soit d'études détaillé, pour le niveau du grade et pour l'emploi dans le secteur public. Pour les équations de 1984 et de 1987, la majorité des variables sont propres à chaque point de référence. Pour l'équation du changement dans les gains, on introduit les variables de contrôle de 1984 ainsi que tout changement qui se produit, dans ces variables de contrôle, entre 1984 et 1987.

Remarquez que le modèle ne comprend pas de variables de contrôle pour la branche d'activité et pour la profession. Le processus de jumelage des diplômés avec les emplois dans les différentes branches d'activité et dans les différentes professions peut dépendre du sexe. L'utilisation de variables de contrôle pour la branche d'activité et pour la profession peut donc masquer un élément de la discrimination dans les gains (Gunderson, 1988). Par contre, on inclut une variable de contrôle pour l'emploi dans le secteur public puisque l'application répandue de programmes visant des groupes cibles et le principe du mérite qui est énoncé pour l'embauchage et la promotion dans ce secteur peuvent créer un ensemble distinct de règles pour le jumelage des emplois.

Résultats

On peut attribuer aux caractéristiques différentes relatives au capital humain et aux antécédents (talents) des hommes et des femmes une partie relativement faible de l'écart entre les gains, en 1984 et en 1987. On peut attribuer à cette proportion 'expliquée' environ un tiers de l'écart entre les gains à chaque point de référence. On peut aussi attribuer aux profils différents de domaines d'études, pour chaque sexe, la majorité de la proportion 'expliquée' de l'écart. L'emploi dans le secteur public était aussi un facteur important, mais son effet allait habituellement dans le sens contraire de celui du domaine d'études -- il tendait à être une caractéristique compensatrice. La différence résiduelle était habituellement assez bien partagée entre les composantes de l'avantage de traitement pour les hommes et du désavantage de traitement pour les femmes.[7]

La moyenne géométrique des gains des hommes dépassait celle des femmes de $3700 en 1984 et de $7000 en 1987 (voir le tableau 3.). Au cours de ces deux années, la différence entre les talents des hommes et ceux des femmes expliquait 35 pour cent de l'écart entre les gains (c.-à-d. $1300 en 1984 et $2500 en 1987). Les répartitions divergentes des domaines d'études étaient les facteurs les plus importants -- constituant 133 pour cent de la différence nette en 1984 et 84 pour cent en 1987.[8] Le pourcentage plus élevé d'hommes détenant une maîtrise était aussi un facteur important en 1984, mais il l'était moins en 1987. L'âge (au cours des deux années) et l'emploi dans le secteur public (en 1984) étaient les facteurs les plus importants pour faire diminuer l'écart expliqué entre les gains. Les femmes dans la population étaient, en moyenne, plus âgées que les hommes, l'âge donnant lieu à des rémunérations positives. De même, un pourcentage plus élevé de femmes travaillaient dans le secteur public au cours des deux années. Cependant, l'emploi dans le secteur public a produit des rémunérations positives en 1984, mais des rémunérations négatives en 1987 -- en bonne partie à cause d'une prime élevée revenant aux hommes qui travaillaient dans le secteur privé en 1987. Le pourcentage plus élevé de femmes dans le secteur public était donc une vari ble compensatrice en 1984, mais il a aidé à accroître l'écart entre les gains en 1987. Dans une veine semblable, les femmes ont obtenu, en 1984, des rémunérations plus élevées pour leur expérience professionnelle antérieure à plein temps (particulièrement si elle était de trois ans ou plus), mais cet effet s'était dissipé en 1987.

Tableau 3. Composantes estimées de l'écart entre les gains, en 1984 et en 1987, des femmes et des hommes qui ont reçu un diplôme universitaire[1] en 1982[2]

	1984		1987	
	$	%	$	%
Caractéristiques différentes	1,300	35	2,500	36
Rémunérations différentes pour les caractéristiques				
- Avantage pour les hommes	1,000	28	2,300	33
- Désavantage pour les femmes	1,400	37	2,200	31
Ecart total entre les gains[3]	3,700	100	7,000	100

[1] Employés à plein temps au cours de chacune des 5 périodes de référence: janvier 1983, octobre 1983, juin 1984, janvier 1986 et mars 1987.

[2] On peut obtenir les résultats détaillés de la décomposition en s'adressant à l'auteur.

[3] Différence entre la moyenne géométrique des gains des hommes et celle des femmes.

Discussion

Les données de l'END et de l'ESD indiquent qu'il existe un écart assez considérable entre les gains des hommes et ceux des femmes qui ont récemment reçu un diplôme universitaire et que cet écart augmente avec le temps. De plus, l'écart persiste parmi presque tous les domaines d'études et tous les niveaux d'études, les détenteurs donnéesobservées pour j variables, b est un vecteur de j coefficients mesurant les rémunérations pour ces variables et u est le terme d'erreur. Des équations identiques des gains sont estimées pour des sous-échantillons d'hommes et de femmes à l'aide des moindres carrés ordinaires (MCO). Une fois les coefficients estimés, le terme d'écart s'élimine, les estimateurs des MCO remplacent b et les indices supérieurs h et f désignent les équations relatives aux hommes et celles qui se rapportent aux femmes respectivement; nous avons donc

$$\ln W^f = \hat{b}^f \; X^f$$

$$\ln W^m = \hat{b}^m \; X^m.$$

Une propriété des estimateurs des MCO est que la somme du produit des coefficients par les moyennes des variables associées correspond à la moyenne de la variable indépendante (celle qui occupe la partie gauche des équations), de sorte que

$$\overline{\ln W}^f = \hat{b}^f \; \overline{X}^f$$

$$\overline{\ln W}^m = \hat{b}^m \; \overline{X}^m.$$

La technique de décomposition est basée sur le fait que la différence entre les gains moyens est une fonction simple des moyennes des variables explicatives et de la rémunération estimée pour ces caractéristiques. Par conséquent, si les hommes et les femmes recevaient la même rémunération pour leurs talents, b^f est égal à b^h, et la différence entre les gains serait attribuable uniquement à des talents différents.

Cotton a reconnu qu'en l'absence de traitement différent, la rémunération des talents se trouverait quelque part entre celle des personnes qui sont favorisées actuellement et celle des personnes qui sont défavorisées actuellement. Il suggère que ces coefficients 'non discriminatoires' soient estimés sous forme de la moyenne pondérée des coefficients pour les hommes et pour les femmes. Donc,

$$\hat{b}^\star = p^m \; \hat{b}^m + p^f \; \hat{b}^f,$$

où $\hat{b}^\star$ est le vecteur de coefficients non discriminatoires et p^m ainsi que p^f sont les proportions de la population totale composées d'hommes et de femmes respectivement.

Après plusieurs étapes simples que nous n'avons pas à reprendre ici, Cotton obtient une décomposition de la différence entre les gains qui comprend trois termes:

$$\begin{aligned}\overline{\ln W}^m - \overline{\ln W}^f &= \hat{b}^\star \; (\overline{X}^m - \overline{X}^f) \\ &+ \overline{X}^m \; (\hat{b}^m - \hat{b}^\star) \\ &+ \overline{X}^f \; (\hat{b}^\star - \hat{b}^f).\end{aligned}$$

Le premier terme représente la composante de l'écart entre les gains attribuable à des talents différents (capital humain et antécédents d'une personne). Le deuxième et le troisième termes divisent la différence résiduelle entre les gains en un avantage de traitement pour les hommes (rémunération des talents supérieure à la moyenne) et en un désavantage de traitement pour les femmes (rémunération des talents inférieure à la moyenne).

La somme des valeurs en dollars exprimées pour ces composantes correspond à la moyenne géométrique des gains des hommes et des femmes, par opposition à la moyenne arithmétique utilisée dans la section précédente. La moyenne géométrique n'est que du logarithme moyen des gains. C'est-à-dire

$$e^{(\overline{\ln W})}.$$

Les équations ont été estimées pour les gains en 1984 et en 1987 ainsi que pour le changement dans les gains entre 1984 et 1987. **La population a été limitée aux diplômés avec des données valides pour les gains en 1984 et en 1987 qui travaillaient à plein temps au cours de chacune des cinq périodes de référence visées par les enquêtes.**[5] Ainsi, la participation des sous-populations des femmes et des hommes est forte et presque égale. Cette définition donne une estimation prudente de l'écart entre les gains puisque plus de femmes que d'hommes ont une expérience de travail qui comporte des interruptions et un plus grand nombre d'entre elles travaillent à temps partiel.[6] On a aussi vérifié la sensibilité des résultats en fonction d'autres définitions de la population et l'on traitera de cette question plus loin.

Les variables indépendantes (explicatives) comprenaient des variables de contrôle pour l'âge, pour la langue, pour la province, pour la mobilité interprovinciale, pour les études postsecondaires effectuées par les parents, pour l'état matrimonial, pour les enfants, pour l'expérience professionnelle avant les études, pour le domaine

Tableau 2. **Rapports entre les gains des femmes et ceux des hommes qui ont reçu, en 1982, un diplômeuniversitaire et qui travaillaient à plein temps en 1984 ou en 1987, selon le domaine d'études et la durée du programme**

Domaine d'études	Durée du programme	Rapport femmes/hommes pour 1984	Rapport femmes/hommes pour 1987
Tous les domaines	1^{er} cycle universitaire	90	83
	Maîtrise/Certificat de 2^e ou de 3^e cycle	85	81
	Doctorat	101	99
Education	1^{er} cycle universitaire	92	89
	Maîtrise/Certificat de 2^e ou de 3^e cycle	83	86
	Doctorat	91	88
Beaux-arts et humanitéss	1^{er} cycle universitaire	99	91
	Maîtrise/Certificat de 2^e ou de 3^e cycle	95	95
	Doctorat	105	94
Commerce, économie et droit	1^{er} cycle universitaire	87	87
	Maîtrise/Certificat de 2^e ou de 3^e cycle	87	89
	Doctorat	..	..
Autres sciences sociales	1^{er} cycle universitaire	94	90
	Maîtrise/Certificat de 2^e ou de 3^e cycle	89	84
	Doctorat	93	91
Sciences agricoles et biologiques	1^{er} cycle universitaire	91	80
	Maîtrise/Certificat de 2^e ou de 3^e cycle	89	84
	Doctorat	87	89
Génie	1^{er} cycle universitaire	91*	89*
	Maîtrise/Certificat de 2^e ou de 3^e cycle	80*	..
	Doctorat	111	119
Sciences médicales et sciences de la santé	1^{er} cycle universitaire	65	54
	Maîtrise/Certificat de 2^e ou de 3^e cycle	77	50
	Doctorat	158*	118*
Mathématiques et sciences physiques	1^{er} cycle universitaire	95	93
	Maîtrise/Certificat de 2^e ou de 3^e cycle	83*	89*
	Doctorat	94*	94*
Moyenne non pondérée		94	92
Moyenne pondérée		87	82

Note: .. taille de l'échantillon trop petite pour que les données soient publiées (coefficients de variation > 25%)

* taille de l'échantillon relativement petite, il faut interpréter les données avec prudence (coefficients de variation compris entre 16.5% et 25%)

Il est important de se rappeler que les résultats de la décomposition sont des estimations sujettes à une erreur de spécification ainsi qu'à une erreur de mesure. Les résultats peuvent être modifiés par des caractéristiques non mesurées du capital humain ou par le libre choix (par ex., le fait qu'un diplômé choisisse une profession plutôt qu'une autre pour des raisons non monétaires). Par conséquent, la décomposition ne peut fournir une preuve directe de discrimination salariale. Par contre, elle peut indiquer les caractéristiques qui pourraient entraîner une récompense différente. Voici une brève description de la méthode de décomposition.

Méthode

La technique de décomposition 'non discriminatoire' exposée par Cotton (1988) est utilisée pour le présent article. Cette technique est une variante d'une méthode qui remonte aux années 1950 et qui a été mentionnée dans des ouvrages portant sur l'économie, sur la sociologie et sur la démographie.[3]

Considérons l'équation des gains suivante:

$$\ln W = b X + u$$

où $\ln W$ est le logarithme naturel des gains annuels[4]; X est une matrice (k,j) de k valeurs de

Tableau 1. Rapports entre les gains des femmes et ceux des hommes pour les diplômés d'université de 1982 employés à plein temps en 1984 ou en 1987, selon le domaine d'études

Domaines d'études	Rapport femmes/hommes pour 1984	Rapport femmes/hommes pour 1987
	%	%
1. Education	87	86
2. Beaux-arts	96	89*
3. Arts appliqués	..	..
4. Journalisme	..	..
5. Autres humanités	98	94
6. Sociologie, anthropologie, démographie	99	97
7. Criminologie	..	..
8. Droit	88	95
9. Economie	88	75
10. Géographie, environnement	83	82
11. Science politique	104	86
12. Psychologie	83	82
13. Autres sciences sociales	90	86*
14. Agriculture	..	..
15. Biochimie, biologie, zoologie	90	95
16. Economie domestique	..	..
17. Médecine vétérinaire	..	..
18. Architecture	..	..
19. Génie	89	89*
20. Foresterie	..	..
21. Architecture paysagiste	..	..
22. Médecine dentaire	..	..
23. Médecine	81	87
24. Sciences infirmières	..	..
25. Optométrie	..	..
26. Pharmacie	..	..
27. Hygiène publique	..	..
28. Informatique	95	91*
29. Mathématiques	97	93*
30. Chimie, géologie, métallurgie		84
31. Météorologie	90	..
32. Physique, autres sciences	..	..
Moyenne non pondérée	89	85
Moyenne pondérée	87	82

Note: .. taille de l'échantillon trop petite pour que les données soient publiées (coefficients de variation > 25%)

* taille de l'échantillon relativement petite, il faut interpréter les données avec prudence (coefficients de variation compris entre 16.5% et 25%)

Décomposition de la différence entre les gains des femmes et ceux des hommes

Dans la section précédente, l'écart entre les gains des diplômés et ceux des diplômées n'a été groupé en catégories qu'en fonction d'une variable ou de deux variables simultanément. Bien que si l'on faisait davantage de classements recoupés ou si l'on utilisait des catégories plus détaillées on pourrait créer des groupes plus comparables, les petites tailles des échantillons dans les cellules limitent considérablement la portée de telles analyses. Par contre, une méthodes multidimensionnelle permet d'étudier simultanément les effets d'un certain nombre de variables et d'évaluer les résultats à l'aide des notes de tests standardisés. Dans la présente section, nous utilisons une technique multidimensionnelle appelée décomposition pour analyser la différence entre les gains en fonction du sexe.

La technique de décomposition est basée sur des régressions linéaires des gains de deux groupes différents; dans le présent cas, les diplômés des deux sexes. Les équations de régression sont structurées selon un modèle du capital humain: les gains sont modélisés comme fonction des études et de l'expérience (investissement dans le capital humain), alors que l'on tient compte des antécédents des personnes ou des caractéristiques démographiques. Comme nous l'avons fait remarquer dans la section précédente, au moins une partie de l'écart entre les gains est attribuable à des différences entre les programmes d'études suivis par les hommes et par les femmes. Il peut aussi en être de même pour l'expérience ou pour n'importe quel élément d'une gamme étendue de caractéristiques relatives aux antécédents d'une personne. La technique de décomposition est principalement un outil servant à estimer la proportion de l'écart entre les gains attribuable aux différences mesurées dans le capital humain et dans les antécédents des hommes et des femmes. La différence qui reste dans les gains est appelée composante résiduelle. Les coefficients de régression permettent de subdiviser la différence résiduelle en rémunérations différentes pour les caractéristiques des personnes.

En dépit de la participation croissante des femmes pour ce qui est du travail à plein temps et de leurs progrès en matière de niveau de scolarité, les gains des femmes sont encore très inférieurs à ceux des hommes. Bien que l'écart entre les gains aie diminué au cours des 20 dernières années, il est encore important. En 1987, les femmes qui ont travaillé à plein temps pendant toute l'année gagnaient, en moyenne, un tiers de moins que leurs homologues masculins. Se peut-il que l'on puisse attribuer un écart aussi considérable à des différences dans la structure par âge, dans les études ou dans l'expérience des travailleurs et des travailleuses? Il serait très difficile de répondre à cette question pour l'ensemble de la population active, mais le panel de l'END et de l'ESD donne un aperçu d'une cohorte récente d'entrants sur le marché du travail dont nous connaissons la majorité des caractéristiques pertinentes. Dans la présente section, nous examinerons l'écart entre les gains parmi des groupes de plus en plus précis de diplômés d'université qui font partie du panel.

Si l'on considère les diplômés d'université de 1982, les femmes travaillant à plein temps en 1984 gagnaient, en moyenne, 24 mille dollars -- ou 87 pour cent du salaire moyen des hommes qui s'élevait à 27 mille dollars. En 1987, le rapport entre les gains des femmes et ceux des hommes avait diminué pour atteindre 82 pour cent, les femmes gagnant, en moyenne, 31 mille dollars comparativement à 38 mille dollars pour les hommes.

L'écart entre les gains était plus petit parmi les diplômés d'université que pour les personnes dans la population active qui ont à peu près le même âge et qui travaillent à plein temps. Parmi l'ensemble des membres de la population active qui ont le même âge que les diplômés d'université en 1982, les gains des femmes s'établissaient, en moyenne, à 18 mille dollars et ceux des hommes à 25 mille dollars -- un rapport de 70 pour cent.[2] De même, en 1987, l'ensemble des femmes dans la population active, pondéré en fonction de l'âge, comparable aux diplômés d'université gagnait 71 pour cent des gains moyens des hommes. La pondération en fonction de l'âge a tendance à diminuer l'écart entre les gains parce que les gains des hommes et ceux des femmes sont plus rapprochés pour les plus jeunes groupes d'âges où l'on retrouve la majorité des diplômés.

Les différences dans les répartitions des domaines d'études des hommes et des femmes sont frappantes. De nombreux domaines d'études tendent à être dominés par un sexe ou l'autre. La répartition différente des domaines d'études peut avoir un effet sur l'écart global entre les gains. Si les hommes tendent à graviter autour des domaines d'études où les récompenses sont le plus élevées, cela pourrait accroître l'écart. Il existe une méthode simple pour vérifier si cet écart existe, il suffit de comparer les rapports entre les gains à l'intérieur des domaines d'études avec le rapport global. Si la moyenne des rapports à l'intérieur des domaines d'études est significativement plus faible que le rapport global, les choix de domaines d'études différents par les hommes et par les femmes expliquent une partie de l'écart entre les gains.

Le rapport entre les gains des femmes et ceux des hommes dans un même domaine d'études était, en moyenne, de 89 pour cent en 1984 et de 85 pour cent en 1987, ce qui entraîne une réduction de 2 et de 3 points respectivement de l'écart global entre les gains.

Bien que l'écart entre les gains soit généralement plus faible à l'intérieur des domaines d'études, les femmes détenant un diplôme dans presque tous les programmes gagnent encore moins que les hommes. En fait, il n'y a qu'un domaine -- la science politique -- où les diplômées gagnaient au moins autant que leurs homologues masculins en 1984. Mais, le pendule des gains avait oscillé dans la direction favorable aux hommes en 1987.

Le niveau du grade est une autre variable qui explique une partie de la stratification des salaires. Les dipLômés d'université détenant un doctorat qui travaillaient à plein temps en 1984 gagnaient 45 pour cent de plus (35 pour cent de plus en 1987) que les personnes détenant un diplôme de premier cycle, les personnes détenant une maîtrise se trouvant entre ces deux groupes. L'écart le plus considérable entre les gains se trouvait au niveau des personnes détenant une maîtrise, les rapports étant de 85 pour cent en 1984 et de 81 pour cent en 1987, comparativement à 90 pour cent et à 83 pour cent respectivement, pour les titulaires d'un diplôme de premier cycle.

L'écart est pratiquement inexistant au niveau du doctorat: les femmes titulaires d'un doctorat gagnaient 1 pour cent de plus que les hommes en 1984 et 1 pour cent de moins en 1987.

La combinaison des effets du domaine d'études et du niveau du grade devrait donc réduire un peu l'écart entre les gains. Et c'est effectivement ce qui se produit. Si l'on calcule la moyenne pour 10 des principaux domaines d'études et pour trois niveaux de grades, les diplômées d'université gagnaient 94 pour cent du salaire de leurs homologues masculins en 1984 et 92 pour cent en 1987. Bien entendu, les détenteurs d'un doctorat, qui ne constituaient qu'une faible partie de la population, ont une influence disproportionnée sur cette moyenne.

De plus, les regroupements de domaines d'études, à ce niveau, présentent quelques problèmes. Par exemple, dans la catégorie "sciences médicales et sciences de la santé" on compare une population masculine composée principalement de médecins à une population féminine où l'on retrouve surtout des diplômées en sciences infirmières. Dans les tableaux croisés, le détail (ou l'homogénéité des groupes) est limité par la variabilité élevée des petites tailles des échantillons à l'intérieur des cellules.

Puisque l'analyse descriptive est limitée à une ou deux variables, on peut concevoir qu'il soit possible d'attribuer aux effets combinés d'autres variables une partie des différences, à l'intérieur des cellules, relatives aux gains des hommes et à ceux des femmes. Par conséquent, dans la section suivante, nous utilisons une technique multidimensionnelle pour étudier les effets simultanés de nombreuses variables sur l'écart entre les gains.

Bien que nous présentions quelques statistiques descriptives et des tableaux croisés, la majeure partie de l'analyse est de nature multidimensionnelle. Une technique nommée "décomposition" est utilisée pour diviser l'écart entre les gains en une composante qui peut être expliquée par les différences dans les caractéristiques de l'instruction et des antécédents ainsi qu'en une composante résiduelle qui pourrait être l'indice de certains désavantages relatifs pour un des groupes étudiés. Les résultats de la décomposition sont présentés pour les gains en 1984 et en 1987.

Les données de l'END et de l'ESD montrent qu'il existe un écart important entre les gains des hommes et ceux des femmes qui ont récemment obtenu un diplôme universitaire. Bien qu'il soit réduit dans certaines catégories, l'écart persiste dans presque tous les domaines et tous les niveaux d'études. Il n'y a qu'environ le tiers de l'écart qui peut être expliqué par des différences dans les caractéristiques de l'instruction et des antécédents des hommes et des femmes. On a aussi trouvé que l'écart entre les gains croissait dans le temps pour la cohorte étudiée.

Un bref commentaire sur les données

L'enquête nationale auprès des diplômés (END), de 1984, et l'enquête de suivi auprès des diplômés (ESD), de 1987, ont permis de recueillir une gamme étendue de renseignements sur les expériences du marché du travail des personnes qui ont reçu, en 1982, un diplôme d'un collège communautaire ou d'une université. Chacune de ces enquêtes comprenait une question où l'on demandait aux répondants d'estimer leurs gains annuels (à mille dollar près) en basant leur évaluation sur l'emploi qu'ils occupaient au moment de l'interview. D'ordinaire, les réponses à cette question ne comprennent que deux chiffres (c.-à-d. 10 à 99 mille dollars). Pour éviter toute représentation de fausse précision, la majorité des données numériques qui figurent dans le présent document sont aussi fournies à deux chiffres significatifs.

L'analyse présentée dans ce rapport ne porte que sur les travailleurs employés à plein temps. Les comparaisons descriptives des gains de 1984 comprennent tous les diplômés travaillant à plein temps en 1984 et une restriction semblable s'applique aux comparaisons des gains de 1987. Puisque le nombre d'heures travaillées varie beaucoup pour les travailleurs à temps partiel et que les enquêtes ne demandaient pas le nombre d'heures travaillées, cette condition permet de nous assurer que nous comparons des quantités de travail approximativement égales. Il s'ensuit aussi que les chiffres sur les gains donnent une approximation des gains reliés à un travail à plein temps pendant toute une année, à cause de la façon dont la question a été posée. Le fait que l'analyse ne porte que sur les personnes travaillant à plein temps donne les tailles d'échantillon maximales ci-après pour les tableaux descriptifs:

	1984	1987
Hommes	5141	4986
Femmes	4032	3689
Total	9173	8675

La taille exacte de l'échantillon pour chacun des tableaux sera un peu plus petite à cause des valeurs manquantes pour les variables étudiées.

Une définition beaucoup plus restrictive a été utilisée pour les analyses multidimensionnelles: seules les personnes qui travaillaient à plein temps lors de cinq périodes de référence distinctes ont été incluses. Le plus grand échantillon résultant est composé de 5971 personnes (3582 hommes et 2389 femmes). Les échantillons d'analyse sont beaucoup plus petits à cause des valeurs qui manquent pour les nombreuses variables incluses dans l'analyse. Dans la mesure du possible, les comparaisons descriptives ont été produites pour la population de régression et vice versa. Aucun des ensembles de résultats n'a été modifié considérablement à la suite de changements apportés aux spécifications de la population.

On peut obtenir des renseignements plus détaillés sur les enquêtes, sous forme de guides de l'utilisateur et de rapports de méthodologie, en s'adressant à la Division des enquêtes-ménages.

Historique

Au cours des 30 dernières années, on a assisté à un changement spectaculaire dans la marché du travail au Canada -- plus particulièrement la participation accrue des femmes dans les emplois à plein temps. Entre 1967 et 1988, la proportion des emplois à plein temps occupés par des femmes a augmenté de 27 pour cent à près de 39 pour cent. Cette proportion devrait continuer de croître puisque les taux d'activité à plein temps des femmes plus jeunes sont plus élevés que dans les cohortes plus âgées.

Les progrès des femmes en matière de niveau de scolarité sont encore plus spectaculaires que leur participation croissante à la population active. Bien que seulement un quart des grades universitaires de premier cycle aient été accordés à des femmes au début des années 1960, ces dernières représentaient plus de la moitié des diplômés à la fin des années 1980. Les femmes sont encore moins nombreuses que les hommes au niveau des études supérieures, mais elles les rattrapent rapidement. La proportion des femmes qui obtiennent un grade de maîtrise a augmenté de 19 pour cent en 1961 à 45 pour cent en 1989. Les femmes ont obtenu moins d'un dixième des doctorats acquis en 1961, mais près du tiers de ces grades en 1989.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

L'ÉCART ENTRE LES GAINS DES HOMMES ET CEUX DES FEMMES PARMI DE RÉCENTS DIPLÔMÉS D'UNIVERSITÉ: LES CINQ PREMIÈRES ANNÉES

T. Wannell[1]

RÉSUMÉ

Le présent rapport se concentre sur la différence entre les gains des hommes et ceux des femmes pour un groupe très fermé -- les diplômés de 1982 des universités qui travaillaient à plein temps en 1984 ou en 1987. Bien que l'écart entre les gains pour cette cohorte jeune et bien instruite soit plus faible que celui qui a été relevé au niveau de la population active dans son ensemble, les diplômées gagnaient moins que leurs homologues masculins dans presque toutes les catégories étudiées. De plus, l'écart a augmenté de 1984 à 1987. Nous avons introduit un modèle multidimensionnel afin de mieux tenir compte des nombreux facteurs qui ont une influence sur l'écart entre les gains. En dépit des variables de contrôle, le modèle ne permettait d'expliquer que le tiers de l'écart entre les gains à chaque point de référence.

INTRODUCTION

Le fait qu'il existe un écart entre les gains des hommes et ceux des femmes ne constitue rien de nouveau. Jusqu'à récemment, les rôles familiaux traditionnels dictaient une division du travail dans le ménage de sorte que la majorité des femmes mariées s'occupaient des travaux ménagers et de la garde des enfants sans être rémunérées. Puisque, dans la majorité des cas, la carrière rémunérée des femmes prenait fin au moment de leur mariage ou de la naissance de leur premier enfant, la majorité des femmes ne pouvaient acquérir les compétences et l'expérience nécessaires pour progresser dans l'échelle des salaires. Il existait donc un grand gouffre entre les salaires des hommes et ceux des femmes dans la population active.

Mais les temps changent. Une combinaison de tendances sociales, démographiques et économiques a entraîné une participation accrue des femmes dans la population active employée à plein temps. Moins de femmes quittent la population active quand elles se marient. Les interruptions pour donner naissance à des enfants ont diminué à cause de la chute à long terme de la fécondité et le coût de ces interruptions est subventionné par l'intermédiaire du régime d'assurance-chômage. A mesure que de plus en plus de femmes faisaient leur entrée sur le marché du travail et qu'elles conservaient des emplois rémunérés à plein temps, l'écart entre les salaires diminuait, cependant, il reste encore important.

En 1987, les femmes qui travaillaient à plein temps toute l'année gagnaient, en moyenne, un tiers de moins que leurs homologues masculins. Bien entendu, un écart aussi important reflète de nombreuses différences entre la main-d'oeuvre masculine et la main-d'oeuvre féminine: structure par âge, études, profession, branche d'activité dans laquelle la personne est employée et expérience acquise. La situation est plus embrouillée parce que ces variables changent dans le temps. La population active actuelle est composée de nombreuses cohortes, chacune possédant un ensemble unique de caractéristiques, qui sont entrées sur le marché du travail dans des conditions très différentes.

Que se passerait-il si nous pouvions suivre une seule cohorte récente d'entrants sur le marché du travail à propos desquels nous connaissons la majorité des caractéristiques importantes relatives au revenu? Trouverions-nous que les hommes et les femmes avec les mêmes titres de compétence recevaient à peu près la même rémunération? Ces questions sont précisément du genre de celles auxquelles l'enquête nationale auprès des diplômés et l'enquête de suivi auprès des diplômés peuvent nous permettre de répondre.

L'enquête nationale auprès des diplômés (END) effectuée en 1984 et l'enquête de suivi auprès des diplômés (ESD) menée en 1987 donnent une perspective unique du statut récent de l'écart entre les gains des femmes et ceux des hommes. La base de sondage utilisée pour ces enquêtes inclut les diplômés de 1982 de toutes les universités au Canada.1 Ces enquêtes ont permis de recueillir une gamme étendue de renseignements relatifs à la démographie, à l'instruction et au marché du travail pour la période allant de 1982 à 1987. A l'aide des données ainsi obtenues, nous avons l'intention, dans le présent article, d'étudier deux questions:

i. compte tenu des augmentations récentes dans l'activité des femmes et dans leur niveau de scolarité, existe-t-il encore un écart entre les gains des hommes et ceux des femmes à compétence égale? Et

ii. quelle est l'évolution de l'écart entre les gains, dans le temps, pour une cohorte particulière?

[1] T. Wannell, Groupe d'analyse des entreprises et du marché du travail, Direction des études analytiques, Statistique Canada, Ottawa (Ontario) Canada. L'auteur désire remercier Monica Boyd, Marie-Claire Couture, Joanne Dubeau, Doug Giddings, Bill Magnus, Yigal Messeri, Debbie O'Dwyer, Garnett Picot et Wendy Pyper.

scolaires des résidents ontariens étudiant à l'extérieur de la province en 1986- 1987 montre que le nombre d'étudiants classés comme cas d'abandon apparent mais qui en fait se seraient inscrits à des établissements à l'extérieur de la province ne dépasse pas 2,300. Ce nombre représente 1.3 % des étudiants ayant été classés à tort comme cas d'abandon apparent qui en fait étaient passés à un établissement à l'extérieur de la province. De plus, puisqu'on a établi que 80 % des étudiants de 1er cycle à temps plein en Ontario déclarent leur NAS et que 4 % des étudiants changent d'établissement, on peut supposer que si 100 % des étudiants déclaraient leur NAS, 5 % des étudiants auraient alors changé d'établissement, ce qui réduit d'encore 1 % le taux apparent d'abandon en cours d'études.

4.4 Trimestre d'automne uniquement

Le SISCU réunit des données sur les effectifs à un seul moment au cours de l'année universitaire, soit en décembre ou en novembre. De 5 % à 10 % de l'ensemble des étudiants ne suivent pas de cours au trimestre d'automne mais en suivent à ceux d'hiver, du printemps ou de l'été. Le SISCU ne tient donc pas compte de ce groupe d'étudiants.

4.5 Cas d'abandon avant la date de dénombrement du SISCU

Les étudiants qui s'inscrivent pour la première fois en septembre et qui abandonnent avant le dénombrement des effectifs du SISCU, qui a lieu en décembre ou novembre, ne figurent pas au SISCU et ne sont donc pas classés comme cas d'abandon.

5. APPLICATIONS À D'AUTRES PROVINCES DU CANADA

La même méthodologie pourrait être appliquée à la Colombie-Britannique, au Manitoba et au Nouveau-Brunswick où l'incidence de la déclaration des NAS est élevée et où la plupart des grades sont déclarés au niveau de l'étudiant (tableau 10). Les autres provinces ont soit la majorité de leurs grades déclarés sous forme agrégée, ce qui empêche l'identification individuelle des étudiants qui obtiennent leur diplôme, soit un faible pourcentage de déclaration de NAS, ce qui ne permet pas d'établir les migrations d'un établissement à un autre.

Les ententes entre les administrations provinciales des Maritimes, selon lesquelles une université de ces provinces peut offrir des programmes spécialisés pour tous les étudiants de la région, ont favorisé les migrations d'un établissement à un autre. Ces migrations sont beaucoup plus fréquents qu'en Ontario. Malheureusement, en raison de la faible incidence de la déclaration des NAS en Nouvelle-Écosse et à l'Île-du-Prince-Édouard, il est impossible de suivre avec précision les mouvements des étudiants entre les établissements des Maritimes. Cette impossibilité de suivre les migrations d'un établissement à un autre entraînerait que bon nombre d'étudiants du Nouveau-Brunswick qui poursuivent leurs études en Nouvelle-Écosse ou à l'Île-du-Prince-Édouard seraient classés comme cas d'abandon.

Tableau 10. Applications aux autres provinces

Province	% de grades et de diplômes de 1er cycle déclarés au niveau de l'étudiant	% d'étudiants à temps plein déclarant leur NAS
Terre-Neuve	--	94
Île-du-Prince-Édouard	82	--
Nouvelle-Écosse	56	33
Nouveau-Brunswick	91	83
Québec	7	27
Ontario	87	80
Manitoba	88	72
Saskatchewan	30	30
Alberta	55	53
Colombie-Britannique	70	92

Note: Les provinces ayant des pourcentages élevés dans les deux colonnes sont de bons éléments pour l'utilisation des fichiers couplés dans le but d'estimer les taux d'abandon en cours d'études et les mouvements des étudiants.

6. CONCLUSION

Le couplage des fichiers du SISCU fournit des renseignements précieux sur les mouvements des étudiants, les taux d'abandon en cours d'études et la provenance des étudiants pour les chercheurs en éducation, les planificateurs en main- d'oeuvre et les administrateurs des universités. En Ontario, le taux d'abandon apparent des étudiants universitaires a diminué au cours de la décennie (1976- 1977 à 1986-1987). Ce facteur a contribué à la croissance continue des effectifs du 1er cycle malgré une diminution de la population du groupe d'âge des 18 à 24 ans. Les taux d'abandon apparent et les mouvements des étudiants peuvent être calculés à partir de n'importe quelle caractéristique des étudiants qui figure au fichier des effectifs du SISCU.

Tableau 8. Migrations d'un établissement à un autre
(pourcentage des étudiants ayant changé d'université d'une année à l'autre)

Type de fréquentation et cycle	1976-77		1979-80		1981-82		1983-84		1986-87	
	a un établissement		a un établissement		a un établissement		a un établissement		a un établissement	
	en Ontario	à l'extérieur de l'Ontario	en Ontario	à l'extérieur de l'Ontario	en Ontario	à l'extérieur de l'Ontario	en Ontario	à l'extérieur de l'Ontario	en Ontario	à l'extérieur de l'Ontario
	(pourcentage)									
Temps plein										
Étudiants de 1er cycle	4.1	0.3	3.5	0.5	4.1	0.5	3.4	0.4	3.4	0.4
Étudiants de 2e et 3e cycles	2.0	0.3	1.6	0.5	2.2	0.7	1.8	0.5	1.9	0.5
Temps partiel										
Étudiants de 1er cycle	2.8	0.1	3.4	0.3	4.2	0.4	3.7	0.2	3.7	0.3
Étudiants de 2e et 3e cycles	1.2	0.1	1.7	0.3	2.0	0.3	1.9	0.2	1.5	0.3

Tableau 9. Migrations d'un établissement à un autre, selon les changements de type de fréquentation et de cycle d'une année à l'autre

Type de fréquentation et cycle Année t	Type de fréquentation et cycle Année t+1	Année t 1976 -77	1979 -80	1981 -82	1983 -84	1986 -87
		(pourcentage)				
Étudiant de 1er cycle à temps plein	Étudiant de 1er cycle à temps plein	4	4	4	4	3
Étudiant de 1er cycle à temps plein	Étudiant de 2e ou 3e cycle à temps plein	31	29	29	30	30
Étudiant de 1er cycle à temps plein	Étudiant de 1er cycle à temps partiel	12	12	13	11	10
Étudiant de 1er cycle à temps plein	Étudiant de 2e ou 3e cycle à temps partiel	30	28	37	30	32
Étudiant de 2e ou 3e cycle à temps plein	Étudiant de 2e ou 3e cycle à temps plein	2	2	3	2	2
Étudiant de 2e ou 3e cycle à temps plein	Étudiant de 2e ou 3e cycle à temps partiel	2	2	1	1	2
Étudiant à la maîtrise à temps plein	Étudiant au doctorat à temps plein	13	13	10	11	11
Étudiant de 1er cycle à temps partiel	Étudiant de 1er cycle à temps plein	14	18	17	12	12
Étudiant de 1er cycle à temps partiel	Étudiant de 2e ou 3e cycle à temps plein	31	37	39	42	39
Étudiant de 1er cycle à temps partiel	Étudiant de 1er cycle à temps partiel	4	6	7	7	6
Étudiant de 1er cycle à temps partiel	Étudiant de 2e ou 3e cycle à temps partiel	31	30	29	30	34
Étudiant de 2e ou 3e cycle à temps partiel	Étudiant de 2e ou 3e cycle à temps plein	7	10	10	9	9
Étudiant de 2e ou 3e cycle à temps partiel	Étudiant de 2e ou 3e cycle à temps partiel	1	1	1	1	1

4. LIMITES

4.1 Étudiants qui quittent le pays

Les étudiants qui quittent le pays pour étudier à l'étranger sont classés à tort dans la catégorie des cas d'abandon. En 1986, environ 17,000 Canadiens étudiaient à l'étranger dans des établissements d'études postsecondaires (universités et collèges communautaires). Le nombre de personnes qui étudiaient à une université canadienne l'année précédente et qui, en conséquence, ont été classées comme cas d'abandon, est inconnu.

4.2 Passages au collège communautaire

Les étudiants qui sont passés à des collèges communautaires ou à des instituts de technologie avant de terminer leurs études universitaires sont aussi classés comme cas d'abandon. Le système d'information sur les collèges de l'Ontario indique qu'environ 1,500 étudiants étaient à l'université l'année précédente. Cela signifie qu'environ 1 % des étudiants de 1er cycle à temps plein sont classés comme cas d'abandon même s'ils poursuivent leurs études dans un collège communautaire.

4.3 Passages à d'autres universités au Canada

On peut établir qu'il y a eu passage d'un établissement à un autre seulement dans les cas où le NAS est déclaré (par exemple, on ne peut établir qu'il y a eu passage entre l'University of Toronto, qui déclare les NAS, et l'Université McGill, qui ne les déclare pas; en conséquence, les étudiants sont classés comme cas d'abandon). Environ 4 % des étudiants changent d'établissement d'une année à l'autre. Les universités de l'Ontario ont un taux de réponse élevé pour les NAS. Cela permet de retrouver les étudiants au sein des établissements ontariens. Toutefois, puisque que les universités québécoises ont un faible taux de réponse pour les NAS, il est peu probable que les passages des établissements de l'Ontario à ceux du Québec puissent être établis; les étudiants risquent ainsi d'être classés à tort comme cas d'abandon apparent. Un examen des antécédents

Tableau 7. Taux de transfert d'une discipline à une autre[1] par principal domaine d'études, cycle et type de fréquentation

	1976-77	1979-80	1981-82	1983-84	1986-87
			(pourcentage)		
Étudiants de 1er cycle à temps plein					
Total	19.5	20.3	20.9	27.1	23.0
Sciences agricoles et biologiques	17.3	18.2	20.0	17.6	17.8
Administration des affaires, gestion et commerce	10.8	14.6	18.5	19.2	15.5
Éducation	10.1	10.6	13.1	11.6	11.6
Génie et sciences appliquées	16.9	17.2	17.5	14.2	16.0
Beaux-arts et arts appliqués	9.7	11.5	12.8	10.6	11.4
Arts et sciences, général	30.7	35.6	33.8	62.1	60.8
Professions de la santé	7.1	7.9	8.4	8.8	9.6
Lettres et sciences humaines	17.0	15.4	16.3	15.9	15.3
Mathématiques et sciences physiques	16.5	15.2	15.3	16.3	15.1
Sciences sociales	13.7	14.0	15.3	13.4	14.0
Étudiants de 2e et 3e cycles à temps plein					
Total	4.9	7.8	5.2	4.6	4.9
Sciences agricoles et biologiques	5.1	11.5	3.7	4.9	3.4
Administration des affaires, gestion et commerce	6.0	4.5	3.5	3.1	2.1
Éducation	9.7	4.1	5.8	3.6	3.8
Génie et sciences appliquées	1.7	3.6	2.3	1.6	1.8
Beaux-arts et arts appliqués	3.0	2.0	3.5	2.2	1.4
Arts et sciences, général	3.7	..	10.3	4.3	5.0
Professions de la santé	7.6	12.3	8.9	6.1	7.6
Lettres et sciences humaines	5.7	4.6	4.3	5.5	3.9
Mathématiques et sciences physiques	2.2	7.9	1.4	2.1	2.1
Sciences sociales	2.8	2.6	5.6	5.4	2.7

[1] Pourcentage des étudiants qui changent de discipline d'une année à l'autre (par exemple de la psychologie à la sociologie, des mathématiques à la physique)

3.4 Provenance des étudiants

Jusqu'à maintenant, nous avons principalement insisté sur le cheminement des étudiants d'une année à l'autre. Changent-ils de discipline, de cycle ou, apparemment, abandonnent-ils leurs études? Il est toutefois possible d'examiner les données sous un autre angle: d'où proviennent les étudiants? S'agit-il de nouveaux étudiants (c.-à-d. d'étudiants non inscrits l'année précédente) ou d'étudiants à temps partiel l'année précédente? Étaient-ils inscrits au même cycle ou à un cycle inférieur? Les fichiers couplés du SISCU peuvent répondre à toutes ces questions. Par exemple, des étudiants de maîtrise à temps plein en Ontario en 1987-1988, 21 % étaient étudiants au 1er cycle l'année précédente, 44 % étaient à la maîtrise, 2 % étaient inscrits à un autre type d'études supérieures et 32 % n'étaient pas inscrits en 1986-1987. Des étudiants à temps plein au doctorat en 1987-1988, 0.5 % étaient étudiants au 1er cycle l'année précédente, 14 % étaient à la maîtrise, 71 % étaient au doctorat et 0.7 % étaient inscrits à un autre type d'études supérieures en 1986-1987.

3.5 Migration d'un établissement à un autre

Il est possible, lorsque l'on dispose des NAS, d'observer les migrations d'un établissement à un autre. En Ontario, où la plupart de ces migrations se produisent entre établissements de la province et où on dispose des NAS de 80 % des effectifs à temps plein et de 85 % des effectifs à temps partiel, on peut suivre la vaste majorité des changements d'établissement. Il est plus difficile d'observer les passages à l'extérieur de l'Ontario, en particulier pour les établissements du Québec, car il est rare que les NAS soient déclarés. Les étudiants qu'on ne peut suivre parce qu'on n'a pas leur NAS sont classés à tort comme cas d'abandon apparent même s'ils poursuivent leurs études dans un autre établissement. Le tableau 8 montre que moins de 4 % des étudiants de 1er cycle à temps plein et 2.4 % des étudiants de 2e et 3e cycles à temps plein ont changé d'établissement entre 1986-1987 et 1987-1988. Les pourcentages ne varient que très peu au cours des dix années de la période de 1976-1977 à 1986-1987.

Les migrations d'un établissement à un autre se produisent le plus souvent lorsque les étudiants modifient leur type de fréquentation (temps plein ou temps partiel) ou encore lorsqu'ils changent de cycle (tableau 9). Par exemple, en 1986-1987, 30 % des étudiants qui passent des études de 1er cycle à temps plein aux études de 2e et 3e cycles à temps plein changent d'établissement contre 3 % pour ceux qui restent comme étudiants de 1er cycle à temps plein et 10 % pour ceux qui passent aux études de 1er cycle à temps partiel. Six pour cent des étudiants de 1er cycle à temps partiel qui restent étudiants de 1er cycle à temps partiel changent d'établissement contre 12 % pour ceux qui passent à temps plein.

Tableau 6. Mouvements des étudiants entre les cycles dans les universités ontariennes

Activité l'année suivante	Total	Type de fréquentation et cycle					
		Non inscrits		Inscrits[1]			
		Cas d'abandon apparent	d'étudiants diplômés[2]	Étudiants de 1er cycle		Étudiants de 2^{e} et 3^{e} cycles	
				Temps plein	Temps partiel	Temps plein	Temps partiel
		Pourcentage du total					
Étudiants de 1er cycle à temps plein							
1976-1977	142,576	15.5	16.0	61.2	5.7	1.6	0.2
1979-1980	135,463	13.4	14.9	64.0	6.1	1.6	0.2
1981-1982	147,365	12.4	13.8	65.1	6.7	1.7	0.2
1983-1984	161,106	12.4	14.3	58.9	6.9	1.2	0.1
1986-1987	164,465	11.7	15.1	64.5	7.0	1.6	0.1
Étudiants de 1er cycle à temps partiel							
1976-1977	59,960	45.5	7.2	5.0	41.3	0.5	0.6
1979-1980	71,615	47.8	7.4	6.3	38.6	0.6	0.6
1981-1982	78,639	47.7	6.5	5.6	38.8	0.6	0.6
1983-1984	85,323	48.3	7.1	7.2	36.1	0.6	0.6
1986-1987	83,726	47.1	7.9	7.4	36.4	0.6	0.6
Étudiants de 2^{e} et 3^{e} cycles à temps plein							
1976-1977	18,298	20.4	16.7	1.6	0.5	51.4	9.4
1979-1980	17,817	19.3	15.0	1.2	0.6	53.9	10.0
1981-1982	19,196	17.0	16.5	1.1	0.8	55.6	9.1
1983-1984	20,636	14.0	17.4	0.9	0.7	58.2	8.8
1986-1987	21,381	13.9	17.7	1.1	0.6	58.3	8.4
Étudiants de 2^{e} et 3^{e} cycles à temps partiel							
1976-1977	12,316	31.2	14.7	0.5	1.2	3.7	48.7
1979-1980	12,041	30.9	14.3	0.4	2.3	4.3	47.9
1981-1982	11,839	27.1	16.7	0.5	4.5	2.3	48.9
1983-1984	11,862	26.6	16.9	0.5	1.9	4.6	49.4
1986-1987	11,455	24.0	19.2	0.4	1.6	5.0	49.8

[1] Comprend les étudiants diplômés qui poursuivent leurs études

[2] Comprend les diplômés qui ne se sont pas réinscrits

Note: Les pourcentages ayant été arrondis, leur somme peut ne pas égaler 100%

3.3 Passage d'une discipline à une autre

Il arrive que les étudiants à l'université changent de majeure plusieurs fois. Peu importe leur type de fréquentation ou leur cycle, la vaste majorité des étudiants demeure dans la même discipline d'une année à l'autre. Le tableau 7 présente les taux de passage d'une discipline à une autre selon le principal domaine d'études (autrement dit il indique le pourcentage d'étudiants qui changent de discipline d'une année à l'autre, par exemple de la psychologie à la sociologie). Le tableau 7 montre qu'entre 1986-1987 et 1987-1988, 23.0 % des étudiants de 1er cycle à temps plein changent de domaine d'études, comparativement à 19.5 % en 1976-1977 et à 27.1 % en 1983-1984. Bon nombre d'étudiants débutent leurs études par le programme général d'arts et sciences et décident un jour d'étudier une discipline précise. C'est pour cette raison que les étudiants du programme général d'arts et sciences sont les plus enclins à changer de discipline.

Le fait que le pourcentage des effectifs du programme général d'arts et sciences ait presque doublé entre 1981-1982 et 1983-1984 s'explique par une modification aux méthodes de codage des disciplines dans une grande université. Avant 1983- 1984, presque tous les étudiants de 1er cycle de cet établissement étaient classés dans la catégorie arts et sciences, général, peu importe leur discipline d'études. Seules quelques disciplines hautement spécialisées étaient codées séparément. En conséquence, on établissait que la plupart des étudiants de 1er cycle demeuraient au programme général d'arts et sciences pour l'ensemble de leurs années d'études de 1er cycle. En 1983-1984, cet établissement a commencé à classer dans des disciplines précises les étudiants qui auparavant étaient classés comme appartenant au programme général d'arts et sciences. Ceux qui demeuraient dans ce programme pouvaient, à n'importe quel moment de leurs études universitaires transférer à une discipline codée séparément, ce qui se répercutait aussi sur le codage des majeures. Depuis que ces transferts sont enregistrés, le taux de transfert a presque doublé, passant de 33.8 % en 1981- 1982 à 62.1 % en 1983-1984. Il s'agit là d'une augmentation artificielle créée par la modification aux méthodes de codage utilisées. Avant 1983-1984, les taux de transfert du programme général d'arts et sciences auraient vraisemblablement été près de ceux d'aujourd'hui, si on avait utilisé la méthode actuelle.

2.5 Décrocheurs reprenant leurs études universitaires

Jusqu'à maintenant, nous avons utilisé les termes de décrocheur ou de cas d'abandon apparent pour décrire les étudiants qui ne se sont pas inscrits l'année suivante et qui n'ont pas obtenu leur diplôme. Avec l'évolution rapide des technologies, de plus en plus de gens considèrent l'éducation comme une activité durable leur permettant de suivre l'expansion rapide des connaissances fondamentales. Un étudiant pourra quitter le système d'éducation puis y revenir plusieurs fois au cours de sa vie. Même si une personne quittant l'université sans avoir obtenu son diplôme est classé comme cas d'abandon apparent, il est possible qu'elle y retourne un jour. Afin de quantifier le nombre d'étudiants classés comme cas d'abandon apparent qui ont repris leurs études, on a couplé les fichiers des effectifs de 1983-1984, 1984-1985 et 1985-1986 et le fichier des grades de 1984. On a relié les enregistrements des étudiants qui se sont inscrits en 1983-1984, mais non en 1984-1985, et qui n'ont pas obtenu leur diplôme en 1984 (autrement dit les cas d'abandon apparent) avec le fichier des effectifs de 1985-1986. Des cas d'abandon apparent des étudiants de 1er cycle à temps plein, 21 % sont retournés aux études en 1985-1986. Les décrocheurs de 1re année sont moins enclins à reprendre leurs études que ceux de 2^{e} et 3^{e} années.

Tableau 5. Étudiants à temps plein en 1983-1984 qui étaient des cas d'abandon apparent et qui ont repris leurs études universitaires en 1985-1986, par cycle et par année d'études

Cycle et année d'études	Nombre de cas d'abandon	% reprenant les études à temps plein	% reprenant les études à temps partiel	% reprenant les études universitaires
Étudiants de 1er cycle				
Total	19,977	14.5	6.5	21.0
1re année	8,756	14.0	4.6	18.6
2^{e} année	5,295	19.5	7.1	26.7
3^{e} année	3,168	12.8	10.4	23.2
4^{e} année	1,515	10.4	8.4	18.8
5^{e} année	24	12.5	8.3	20.8
Sans objet	1,182	6.3	4.9	11.3
Étudiants de 2^{e} et 3^{e} cycles				
Total	2,889	7.2	4.6	11.8
Maîtrise	1,081	6.9	9.0	15.9
Doctorat	472	9.3	4.9	14.2
Diplôme de 2^{e} et 3^{e} cycles	46	8,7	2.2	10.9
Autres	96	6.3	6.3	12.5

3. MOUVEMENTS DES ÉTUDIANTS

3.1 Introduction

Le couplage permet de repérer non seulement les étudiants qui ne se trouvent pas au fichier l'année suivante, soit les cas d'abandon, mais aussi de comparer les caractéristiques des étudiants d'une année à l'autre pour ceux qui s'y trouvent. Par exemple, les étudiants peuvent changer de cycle, de discipline, d'année d'études, d'établissement ou passer des études à temps plein à des études à temps partiel, et inversement. Les fichiers couplés du SISCU permettent essentiellement de suivre les étudiants ou de mesurer les mouvements des étudiants, qui découlent d'un changement de situation scolaire.

3.2 Passage d'un cycle à un autre

Avec la baisse des taux d'abandon apparent pour les étudiants de 1er cycle à temps plein, on observe des pourcentages légèrement plus élevés en 1986-1987 qu'en 1976-1977 d'étudiants qui demeurent à temps plein au 1er cycle ou qui passent à temps partiel aux études de 1er cycle. Les étudiants de 1er cycle à temps plein en 1986-1987 ne semblent pas plus enclins que la cohorte de 1976- 1977 à passer aux études de 2^{e} et 3^{e} cycles à temps plein; par contre, ils sont moins susceptibles de poursuivre des études de cycle supérieur à temps partiel. La cohorte d'étudiants de 1er cycle à temps partiel de 1986-1987 est moins portée à demeurer aux études à temps partiel et plus encline à passer aux études à temps plein que la cohorte de 1976-1977. Les étudiants de 2^{e} et 3^{e} cycles à temps plein de 1986-1987 sont plus susceptibles de demeurer aux études supérieures à temps plein et moins enclins à passer aux études à temps partiel que ceux de 1976-1977.

En 1986-1987, plusieurs disciplines particulières des étudiants de 1er cycle à temps plein ont des taux d'abandon apparent beaucoup plus élevés que la moyenne. Il s'agit de la bibliothéconomie (26.5 %), de l'agriculture (23.3 %), de la santé publique (20.3 %) et des arts de la scène (sauf la musique) (19.7 %). La bibliothéconomie et les arts de la scène ont des taux d'abandon élevés depuis 1976-1977 alors que l'agriculture et la santé publique avaient des taux moyens ou des taux en-dessous de la moyenne en 1976-1977.

2.4 Cas d'abandon par année d'études

Le tableau 3 montre que les étudiants de première année ont, au cours de la dernière décennie, toujours eu les taux d'abandon apparent les plus élevés. Tous les taux d'abandon apparent de chaque année d'études ont diminué depuis 1976- 1977.

Tableau 3.
Taux d'abandon apparent des étudiants à temps plein aux programmes de baccalauréat d'une durée de trois ou quatre ans, par année d'études

Année d'études	1976-77	1978-80	1981-82	1983-84	1986-87
			(pourcentage)		
1re année	21.6	18.1	16.9	16.3	15.6
2^{e} année	15.1	12.8	11.1	11.4	11.4
3^{e} année	10.4	9.4	8.8	8.6	8.4
4^{e} année	9.0	8.2	7.5	8.3	7.1

En 1976-1977, 21.6 % des étudiants de 1re année en 1976-1977 sont classés comme cas d'abandon apparent, ne se sont pas inscrits en 1977-1978 et n'ont pas reçu leur diplôme. Il serait peut-être plus intéressant de connaitre le pourcentage de la cohorte de nouveaux étudiants qui abandonnent sans obtenir un diplôme. Ce pourcentage peut être obtenu en couplant les fichiers des effectifs et ceux des grades sur plusieurs années afin de suivre une cohorte de nouveaux étudiants au cours de leur cheminement à l'université. Toutefois, ce couplage coûterait cher, exigerait beaucoup de temps et, en outre, les taux ne seraient disponibles que quatre à cinq ans (période suffisante pour l'obtention d'un diplôme) après l'entrée des étudiants à l'université. On dispose cependant d'une alternative: on peut estimer les taux d'abandon d'une cohorte en appliquant les taux actuels d'abandon apparent et d'obtention d'un diplôme à une cohorte fictive de nouveaux étudiants selon la méthode suivante.
Soit:

E_t^i = le nombre d'étudiants de l'année t entrant à l'année d'études i

d_t^i = le taux d'abandon apparent pour l'année t de l'année d'études i (c.-à-d. les taux du tableau 3)

g_t^i = le taux d'obtention d'un diplôme de l'année d'études i pour l'année t (soit le pourcentage)

D_t = le pourcentage estimé de la cohorte de nouveaux entrants à l'université qui abandonnent avant l'obtention de leur diplôme

Donc, pour toutes les années d'études i,

$$E_t^{i+1} = E_t^i \ (1-d_t^i-g_t^i)$$

$$D_t = \Sigma_i \ d_t^i E_t^i / E_t^i * 100\%$$

Grâce à cette méthode, et aux taux d'abandon apparent du tableau 3, on peut calculer le pourcentage apparent des cas d'une cohorte d'étudiants abandonnant l'université sans détenir de diplôme. Le tableau 4 montre que 43.9 % des nouveaux étudiants de 1976-1977 auraient abandonné contre 34.3 % pour la cohorte de 1986-1987.

Tableau 4.
Taux d'abandon[1] d'une cohorte d'étudiants à un baccalauréat d'une durée de trois ou quatre ans

1976-1977	43.9%
1979-1980	41.1%
1981-1982	37.4%
1983-1984	35.8%
1986-1987	34.3%

[1] Pourcentage de la cohorte de nouveaux entrants à l'université qui n'obtiennent pas leur diplôme

2.2 Cas d'abandon par cycle

La plupart des étudiants de 1er cycle sont inscrits à un programme qui mène à un baccalauréat, normalement après trois ou quatre années d'études. En 1986-1987, moins de 4,000 étudiants de 1er cycle à temps plein sur 180,000 suivent des cours ne menant pas à un diplôme. On les classe dans la catégorie "autres" des étudiants de 1er cycle. C'est dans ce groupe que l'on retrouve le plus haut taux d'abandon apparent pour l'ensemble des étudiants de 1er cycle à temps plein alors que les étudiants inscrits aux programmes de premier cycle professionnel (soit M.D., D.D.S., D.V.M., L.L.B.) ont les taux d'abandon apparent les moins élevés.

Le taux d'abandon apparent des étudiants de 1er cycle à temps partiel demeure relativement stable entre 1976-1977 et 1986-1987 (de 45 % à 48 %). Le taux d'abandon apparent des étudiants de 2e et 3e cycles à temps partiel passe de 31 % en 1976-1977 à 24 % en 1986-1987. Bon nombre d'étudiants à temps partiel classés comme cas d'abandon apparent pourront éventuellement retourner aux études.

Tableau 1. Taux d'abandon apparent[1] dans les universités ontariennes par type de fréquentation et par cycle

Type de fréquentation et cycle	1976-77	1979-80	1981-82	1983-84	1986-87
			(pourcentage)		
Temps plein					
Étudiants de 1er cycle	15.5	13.4	12.4	12.4	11.7
Baccalauréat	15.1	13.1	11.9	11.8	11.3
Premier cycle professionnel (M.D., D.D.S., D.V.M., L.L.B.)	4.4	3.3	3.0	2.8	2.9
Diplôme de 1er cycle	22.5	18.1	16.6	17.9	13.1
Autres	46.7	45.0	43.9	42.4	39.4
Étudiants de 2e et 3e cycles	20.4	19.3	17.0	14.0	13.9
Maîtrise	16.7	15.9	13.1	9.7	10.0
Doctorat	12.9	11.8	10.6	8.3	8.3
Diplôme de 2e et 3e cycles	26.5	23.7	20.1	15.5	10.7
Autres	23.0	36.5	30.4	25.4	29.6
Temps partiel					
Étudiants de 1er cycle	45.5	47.8	47.7	48.3	47.1
Baccalauréat	37.4	35.1	35.9	36.7	36.1
Diplôme de 1er cycle	42.2	53.1	52.4	51.0	54.3
Autres	67.9	66.1	64.3	65.4	64.5
Étudiants de 2e et 3e cycles	31.2	30.9	27.1	26.6	24.0
Maitrise	27.2	28.0	22.8	22.7	20.0
Doctorat	27.5	28.7	25.1	22.8	21.8
Diplôme de 2e et 3e cycles	47.2	36.8	42.0	36.9	39.0
Autres	55.1	56.8	57.9	57.0	61.0

[1] Les taux d'abandon apparent expriment le pourcentage d'étudiants qui figurent au fichier du SISCU une année donnée mais pas l'année suivante et qui n'ont pas obtenu un diplôme. Il représente le pourcentage d'étudiants qui apparemment ont abandonné ou encore ont cessé leurs études dans une université canadienne sans avoir obtenu leur diplôme. L'étudiant dont les fichiers ne peuvent être couplés est classé comme décrocheur ou cas d'abandon apparent.

2.3 Cas d'abandon par domaine d'études

Les taux d'abandon apparent varient considérablement selon le domaine d'études des étudiants, comme l'indique le tableau 2. Les étudiants du programme général d'arts et sciences ont les taux d'abandon apparent les plus élevés pour l'ensemble de la période, suivis par les étudiants en beaux-arts et arts appliqués.

Tableau 2. Taux d'abandon apparent des étudiants de 1er cycle à temps plein selon le principal domaine d'études

Principal domaine d'études	1976-77	1979-80	1981-82	1983-84	1986-87
			(pourcentage)		
Total	15.5	13.4	12.4	12.4	11.7
Sciences agricoles et biologiques	13.0	12.8	13.5	13.6	10.5
Administration des affaires, gestion et commerce	16.7	10.6	9.7	9.1	9.4
Éducation	11.4	9.9	8.9	9.3	7.8
Génie et sciences appliquées	15.0	11.7	10.3	9.2	8.7
Beaux-arts et arts appliqués	18.1	17.5	16.0	16.7	14.6
Arts et sciences, général	20.4	18.1	16.9	16.9	18.7
Professions de la santé	4.5	4.4	3.2	3.2	3.1
Lettres et sciences humaines	16.3	15.6	14.6	13.8	11.5
Mathématiques et sciences physiques	11.9	11.5	9.8	10.5	10.5
Sciences sociales	12.8	12.4	12.2	12.4	11.5

1979-80, 1981-82, 1983-84 et 1986-87. Chacun de ces fichiers d'effectifs du SISCU a été couplé avec le fichier du SISCU de l'année académique suivante et avec le fichier des grades pour la même année. (Ex: le fichier des effectifs de 1986-87 a été couplé au fichier des effectifs de 1987-88 et au fichier des grades de 1987).

2. ABANDON EN COURS D'ÉTUDES

2.1 Introduction

On appelle cas d'abandon apparent ou décrocheur apparent, l'étudiant dont le nom, d'une part, paraît au fichier couplé la première année mais non la deuxième et, d'autre part, ne figure pas sur la liste des diplômés. On a recours au terme "apparent" parce que la technique de couplage n'est pas parfaite: certains étudiants qui poursuivent leurs études ne sont pas couplés parce qu'ils ont changé d'établissement et qu'on n'a pu les retrouver dans leur nouveau lieu d'études. On les a classés à tort comme cas d'abandon. Dans bon nombre de cas, les numéros d'assurance sociale figurent au SISCU, ce qui permet de suivre les étudiants qui ont changé d'établissement. On dispose, pour 1986-1987, des NAS de 80 % des étudiants à temps plein et de 85 % des étudiants à temps partiel de l'Ontario. Sont incorrectement classés comme cas d'abandon les étudiants dont le NAS ne figure pas au fichier et qui changent d'établissement.

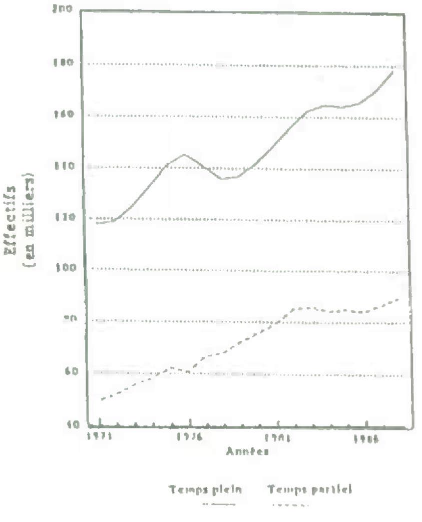
Le taux d'abandon apparent représente le pourcentage d'étudiants qui abandonnent entre les dates de dénombrement des effectifs, soit le 1er novembre d'une année et le 1er novembre de l'année suivante. Il n'exprime pas le taux d'échec d'une cohorte d'étudiants au cours de leur formation universitaire. Une cohorte de nouveaux entrants à l'université peuvent abandonner à n'importe quel moment au cours des trois, quatre années ou plus qu'ils passent à l'université. Le taux d'abandon apparent reflète le pourcentage de décrocheurs qui quittent l'université sans obtenir leur diplôme au cours de l'année.

En Ontario, les effectifs du 1er cycle augmentent, comme l'indique le graphique 1. La croissance observée depuis le début des années 80 s'est produite malgré une baisse de la taille de la population d'où provient traditionnellement les effectifs des universités, soit le groupe d'âge des 18 à 24 ans (graphique 2). La hausse des effectifs est le résultat:

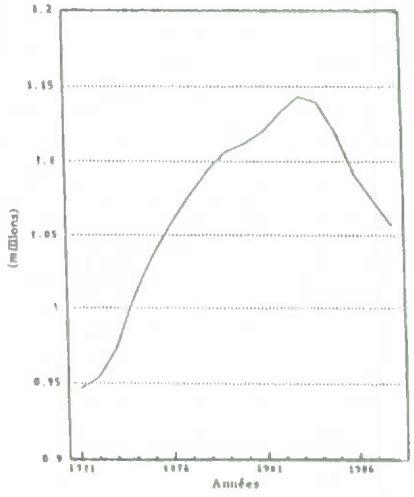
- d'un pourcentage plus élevé de la population d'étudiants complétant leurs études secondaires. En Ontario, le nombre de diplômés de 13^{e} année en rapport à la population des 18 ans passe de 28.8 % en 1970-1971 à 34.5 % en 1986-1987.
- d'une hausse du pourcentage des diplômés de 13^{e} année poursuivant immédiatement leur études universitaires à temps plein. Le pourcentage passe de 65.6 % en 1976-1977 à 73.1 % en 1987-1988.
- d'un accroissement du nombre de personnes de plus de 24 ans inscrites aux études universitaires. Entre 1971-1972 et 1988-1989, les effectifs à temps plein augmentent de 75 % chez les personnes de plus de 24 ans et de 45 % pour le groupe des 18 à 24 ans. Malgré l'importante hausse des effectifs des plus de 24 ans, ceux-ci ne représentent tout de même que 19.6 % des effectifs étudiants à temps plein en 1988-1989.
- d'une diminution du taux d'abandon dans les universités ontariennes.

Entre 1976-1977 et 1986-1987, le taux d'abandon apparent pour les étudiants de 1er cycle passe de 15.5 % à 11.7 %. Ces chiffres représentent le pourcentage d'étudiants de 1er cycle à temps plein qui ne figurent pas au fichier des effectifs l'année suivante et qui n'ont pas obtenu leur diplôme. Les taux d'abandon apparent des étudiants de 2^{e} et 3^{e} cycles à temps plein enregistrent aussi une baisse, passant de 20.4 % en 1976-1977 à 13.9 % en 1986-1987.

Graphique 1. Effectifs du 1er cycle en Ontario 1971-1988



Graphique 2. Population de l'Ontario âgée de 18 à 24 ans 1971-1988



Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MOUVEMENTS DES ÉTUDIANTS DANS LES UNIVERSITÉS DE L'ONTARIO

W. Clark[1]

RÉSUMÉ

Statistique Canada réunit des données sur tous les étudiants à l'université inscrits en décembre de l'année académique. Quatre-vingt-dix-neuf pour cent des données sur les effectifs ont comme source les enregistrements, au niveau de l'étudiant, du système d'information statistique sur la clientèle universitaire (SISCU). On peut estimer les mouvements des étudiants, qui illustrent le cheminement des étudiants depuis leur entrée à l'université jusqu'à l'obtention de leur diplôme, en rattachant les fichiers des effectifs et les fichiers des grades pour plusieurs années consécutives. Le couplage des fichiers du SISCU permet aussi de quantifier les transferts des étudiants d'un établissement à un autre, les changements de discipline et les taux d'abandon en cours d'études. Une série chronologique des mouvements des étudiants dans les universités ontariennes est présentée.

1. INTRODUCTION

Le système d'information statistique sur la clientèle universitaire (SISCU) est une base de données permettant d'obtenir des statistiques sur les effectifs (inscriptions) et les diplômes universitaires à la grandeur du Canada. Les fichiers remontent à l'année universitaire 1972-1973 pour les effectifs et à 1970 pour les grades. On recueille les données auprès des régistraires des universités partout au pays. Les éléments d'information comprennent des renseignements d'intérêt éducationnel et universitaire et un vaste éventail de caractéristiques essentielles sur chaque étudiant, comme le sexe, l'âge, le statut d'immigrant, le pays de citoyenneté, l'activité scolaire de l'étudiant l'année précédente, le cycle, la durée du programme et le nombre de crédits auxquels il est inscrit, la spécialité, etc. Sont inclus tous les étudiants qui sont inscrits à des cours donnant droit à l'obtention de crédits, et qui les suivent, dans le cadre d'un programme menant à un diplôme ou un certificat dans un établissement en décernant. (Sont aussi inclus les étudiants ne cherchant pas à obtenir un diplôme ou un certificat mais qui suivent des cours donnant droit à l'obtention de crédit.) Quatre-vingt-dix-neuf pour cent des données sur les effectifs et 65 % des données sur les grades ont comme source des enregistrements au niveau de l'étudiant.

Le SISCU donne un instantané des effectifs au 1^er^ décembre ou, pour l'Ontario, au 1^er^ novembre. Les données stockées dans le SISCU sur les universités ontariennes sont en majeure partie déclarées au niveau de l'étudiant, tant pour les effectifs que les grades, ce qui permet le couplage des fichiers des effectifs et des fichiers des grades. Seuls le Royal Military College, le Ontario Bible College, le Ontario Theological Seminary et le Collège Dominicain ne sont pas intégrés aux données présentées ici car ces établissements déclarent leurs effectifs sous forme agrégée pour certaines années ou pour toutes les années à l'étude. Ces établissements représentent 0.7 % des effectifs des universités ontariennes en 1988-1989.

La plupart des renseignements du SISCU sont réunis pour les différents étudiants grâce au numéro d'idendité unique donné à chaque étudiant par l'établissement et, dans bon nombre de cas, à l'aide du numéro d'assurance social (NAS), qui est propres à l'individu pour tout le pays. Il est possible de coupler les enregistrements des étudiants d'une année à l'autre à l'aide de ces identificateurs exclusifs. On peut alors calculer les mouvements des étudiants entre une période donnée et la période suivante. De plus, comme bon nombre d'universités ontariennes déclarent pour chacun des étudiants les données relatives aux diplômes décernés, on est en mesure de relier les fichiers couplés des effectifs avec les fichiers des grades. En ce qui a trait aux établissements qui ne déclarent pas les grades au niveau de l'étudiant - soit Queen's University (pour certaines années), Carleton (pour certaines années), Trent et Laurentian, on impute les couplages des grades avec les effectifs selon le nombre de diplômes accordés par cycle et par domaine d'études et selon les taux d'obtention d'un diplôme pour les autres années ou encore pour les autres établissements. Ainsi, on calcule non seulement les mouvements des étudiants en pourcentage mais aussi les taux d'obtention d'un diplôme.

Le couplage des fichiers administratifs du SISCU permet de comparer les caractéristiques des étudiants d'une année à l'autre et, en conséquence, de calculer les mouvements des étudiants. Combien changent de discipline? Combien passent des études à temps plein aux études à temps partiel? Combien changent d'établissement? Combien ne figurent plus au fichier du SISCU (autrement dit abandonnent)? Combien passent d'un programme spécialisé à un programme général ou encore ne réussissent pas à passer à l'année suivante? On peut répondre à toutes ces questions grâce aux fichiers couplés.

Des données couplées pour cinq années choisies pour les universités de l'Ontario, couvrant des années académiques entre 1976-77 et 1986-87, sont présentées dans cet article. Les cinq années choisies sont 1976-77,

[1] W. Clark, Division de l'éducation, de la culture et du tourisme, 16^e^ étage, Immeuble R.-H. Coats, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6.

SECTION 8

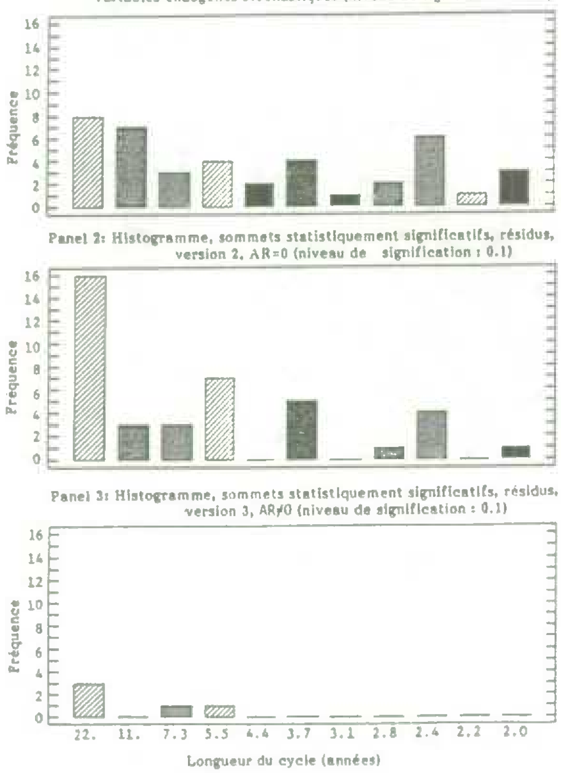
ÉDUCATION

Tableau 2: Sensibilité de la période de certaines racines, version 3

(1) Variable	(2) Sensibilité	(3) Variable	(4) Sensibilité	(5) Variable	(6) Sensibilité	(7) Variable	(8) Sensibilité
				RACINE 3			
KAP	-1.33			KAP(-2)	2.70	KAP(-3)	-1.21
RS	3.87	RS(-1)	-9.39	RS(-2)	-2.95		
				RACINE 12			
DNPE	-444.27	DNPE(-1)	1229.07	DNPE(-2)	-1128.83	DNPE(-3)	343.91
				RACINE 15			
DNNE	133.26	DNNE(-1)	-486.09	DNNE(-2)	566.39	DNNE(-3)	-213.56
DNTFPE	-392.91	DNTFPE(-1)	1128.56	DNTFPE(-2)	-1072.38	DNTFPE(-3)	336.73
				RACINE 16			
DNNE	-392.92	DNNE(-1)	1128.59	DNNE(-2)	-1072.41	DNNE(-3)	336.74
DNTFPE	133.28	DNTFPE(-1)	-486.16	DNTFPE(-2)	566.45	DNTFPE(-3)	-213.58
				RACINE 18			
RS	-4.79	RS(-1)	7.63	RS(-2)	-2.80		
DNPE	-12.02	DNPE(-1)	34.20	DNPE(-2)	-31.33	DNPE(-3)	9.18
KAP	-4.30	KAP(-1)	11.13	KAP(-1)	-9.10	KAP(-3)	2.32
NIC	-1.12	NIC(-1)	0.89	NIC(-2)	0.87		

Nota: Les variables indiquées dans la colonne (1) font partie du membre de gauche des équations qu'elles représentent dans la version 3. Les variables des colonnes (3), (5) et (7), qui sont essentiellement des variables endogènes décalées, font partie du membre de droite des mêmes équations. Seules les valeurs de sensibilité les plus élevées figurent dans ce tableau.

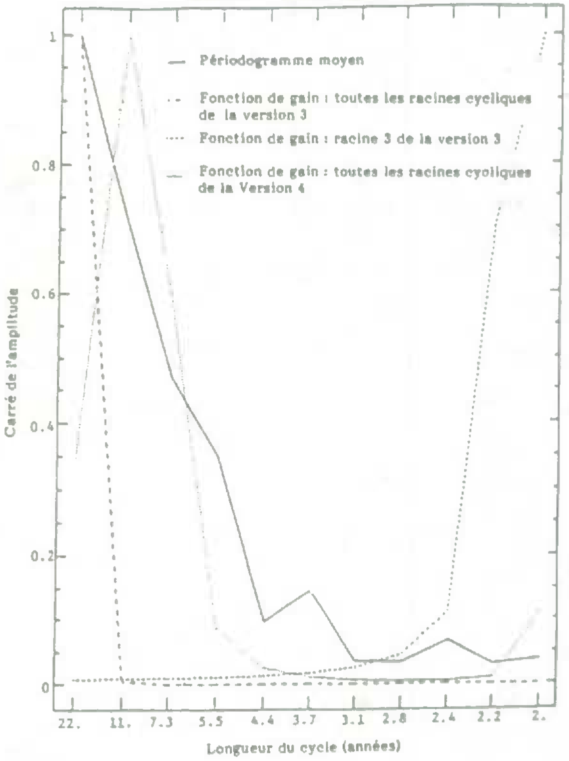
Figure 1: Panel 1: Histogramme, sommets statistiquement significatifs, variables endogènes stochastiques (niveau de signification : 0.1)



Panel 2: Histogramme, sommets statistiquement significatifs, résidus, version 2, AR=0 (niveau de signification : 0.1)

Panel 3: Histogramme, sommets statistiquement significatifs, résidus, version 3, AR≠0 (niveau de signification : 0.1)

Figure 2 : Périodogramme moyen pour les variables stochastiques et fonctions de gain pour les versions 3 et 4



Burmeister, E. and L. Klein, (1974), "Symposium Econometric Model Performance: Comparative Simulation Studies of Models of the U.S. Economy", *International Economic Review*, June 1974, Oct. 1974 and Feb. 1975.

deBever, L., Foot, J.F. Helliwell, G.V. Jump, T. Maxwell, J.A. Sawyer and H.E. Waslander, (1979), "Dynamic Properties of Four Canadian Macro-Models: a Collaborative Research Project", *Canadian Journal of Economics*, 12(2), 133-194.

Duguay, P. and Y. Rabeau, (1987), *Les Effets Macro-Economiques des Déficits Budgétaires: Résultats d'un Modèle de Simulation*, Rapport Technique 47, Ottawa, Banque du Canada, novembre.

Duguay, P. and Y. Rabeau, (1988), "A Simulation Model of Macro-Economic Effects of Deficit", *Journal of Macroeconomics*, Vol. 10, Number 4, Fall, 539-564.

Evans, M.K., L.R. Klein and M. Saito, (1972), "Short-Run Prediction and Long-Run Simulation of the Wharton Model" in B. Hickman, *Econometric Models of Cyclinal Behavior*, Vol. 1, 137-200.

Fromm, G., L.R. Klein and G. Schink, (1972), "Short and Long-Term Simulations with the Brookings Model" in *Econometric Models of Cyclinal Behavior*, Vol. 1, 201-310.

Green, G., M. Liebenberg and A. Hirsch, (1972), "Short and Long-Term Simulations with the OBE Econometric Model" in *Econometric Models of Cyclinal Behavior*, Vol. 1, 25-136.

Hickman, B. (editor), (1972), *Econometric Models of Cyclinal Behavior, Vol. I and II*, New York: National Bureau of Economic Research, Columbia University Press.

Howrey, E.P., (1972), Dynamic Properties of a Condensed Version of the Wharton Model" in *Econometric Models of Cyclical Behavior*, Vol. 2, 601-671.

Kuh, E., J.W. Neese and P. Hollinger, (1985), *Structural Sensitivity in Econometric Models*, New York: John Wiley & Sons.

O'Reilly, B., G. Paulin and P. Smith, (1983), *Responses of Various Econometric Models to Selected Economic Policy Shocks*, Technical Report No. 38, Ottawa: Bank of Canada.

Priestley, M.B., (1981), *Spectral Analysis of Time Series, Vol. 1: Univariate Series*, London: Academic Press.

Troll (1983), *Program LIMO*, Technical Report No. 34, Massachusetts Institute of Technology, Cambridge, Mass., December.

Tableau 1: Données sur les racines caractéristiques

Racines caractéristiques (numéro)	Version 1		Version 2		Version 3		Version 4	
	Valeur absolue de la racine	Périodicité (années)	Valeur absolue de la racine	Périodicité (années)	Valeur absolue de la rachine	Périodicité (années)	Valeur absolue de la rachine	Périodicité (années)
1	1.13		1.02		1.07		1.07	
2	1.03		1.02	23.43	1.03		1.03	
3	1.03	17.2	1.01		1.01	2.08	1.01	2.02
4	1.02		1.01		1.01		1.01	
5	1.01		0.99		1.01		1.01	
6	1.00		0.99		1.00		1.00	
7	1.00	92.17	0.99	89.86	1.0		1.00	
8	1.00		0.96	15.88	1.0		1.00	
9	0.99		0.96	15.88	1.0	67.48	1.00	67.43
10	0.96	15.89	0.94	89.14	1.0		1.00	
11	0.96	15.88	0.92	14.24	1.0		1.00	
12	0.94	14.20	0.92		0.99	19.78	0.99	
13	0.93	55.56	0.92	180.56	0.99		0.97	
14	0.93		0.73	21.60	0.97		0.96	8.89
15	0.92		0.03	3.48	0.96	15.89	0.96	
16	0.92				0.96	15.89	0.94	
17	0.83				0.96		0.91	
18	0.00				0.95	8.51	0.86	
25					0.87	21.96		
29							0.58	11.94
34					0.58	11.94		

Nota: Seules les racines caractéristiques complexes présentent des cycles dont la périodicité est exprimée en années (voir note explicative 3).

la période-échantillon. Cependant, l'effet de ces transformations sur les propriétés cycliques des modèles n'a pas été analysé.

3. Si $R_h = a_h + b_h.i$ définit la racine complexe h, a et b étant respectivement la partie réelle et la partie imaginaire de la racine et $i = \surd -1$, alors la périodicité de cette racine est définie par l'expression $2\pi/(\mathrm{Arctan}\,(b_h/a_h))$.

4. Dans leurs simulations à long terme, au lieu d'utiliser des équations transformées comme l'équation (5), Fromm et coll. (1972), Green et coll. (1972) et Evans et coll. (1972) ont multiplié les constantes de leurs équations par un facteur qui exprime l'autocorrélation de premier ordre des résidus des équations.

5. Étant donné T observations $X_1, \ldots, X_T$, la fonction $P(f)$, appelée périodogramme, est définie pour toutes les valeurs de f (fréquence) dans l'intervalle $-\pi \leq f \leq \pi$ par l'équation

$$P(f) = \frac{1}{2\pi} \left| \sum_{t=0}^{T-1} X_t \, e^{-i2\pi ft} \right|^2$$

où $P(f)$ est évaluée aux fréquences $0,\ 2\pi/t,\ 4\pi/t,\ \ldots$. Pour la construction du périodogramme, on suppose que la série est stationnaire. Pour ne pas éliminer les cycles qui auraient une très longue périodicité, on a extrait la tendance des variables à l'aide d'une formule de tendance temporelle linéaire simple.

6. On peut imaginer que chacune des racines caractéristiques du modèle se rattache à un polynôme $g(L)$ en L (opérateur de décalage) dans le domaine temporel. La réponse d'une variable endogène quelconque Y_t du modèle à une variable aléatoire X_t est définie par l'équation

$$Y_t = \sum_{-\infty}^{\infty} g(L)\ X(t-L) \qquad (a)$$

La valeur absolue et la périodicité de chaque racine caractéristique nous renseignent sur la forme de $g(L)$. En appliquant la transformée de Fourier à (a) et en calculant l'espérance mathématique, nous obtenons l'équation (b), qui définit la représentation spectrale de la fonction de réponse (c'est-à-dire la fonction de gain) dans le domaine des fréquences.

$$\Gamma_{YY}(f) = T(f).\Gamma_{XX}(f) \qquad (b)$$

où Γ_{YY} et Γ_{XX} sont les fonctions de densité spectrale et $T(f)$ est définie comme ci-dessous

$$T(f) = \sum_{-\infty}^{\infty} g(L)\ e^{-ifL} \qquad (c)$$

D'après l'équation (b), il est facile de voir que selon que $T(f)$ sera supérieure, égale ou inférieure à 1, l'amplitude du cycle à n'importe quelle fréquence f en Y sera supérieure, égale ou inférieure à l'amplitude du cycle à la fréquence correspondance en X. Priestley (1981, p. 273) montre que la fonction de gain du modèle est égale au produit des fonctions de gain qui se rapportent respectivement aux racines caractéristiques du modèle.

7. La sensibilité de la période et de la valeur absolue d'une racine peut être calculée à l'aide des matrices D ou E du système linéarisé par les formules d'élasticité suivantes :

	PÉRIODE	VALEUR ABSOLUE
Matrice D	$\frac{\partial\lvert t_h\rvert}{\partial d_{ij}} \cdot \frac{d_{ij}}{\lvert t_h\rvert}$	$\frac{\partial\lvert R_h\rvert}{\partial d_{ij}} \cdot \frac{d_{ij}}{\lvert R_h\rvert}$
Matrice E	$\frac{\partial t_h}{\partial e_{ij}} \cdot \frac{e_{ij}}{\lvert t_h\rvert}$	$\frac{\partial R_h}{\partial e_{ij}} \cdot \frac{e_{ij}}{\lvert R_h\rvert}$

où R_h est le module de la racine caractéristique h, d_{ij} et e_{ij} sont les valeurs des éléments i, j des matrices D et E respectivement, et t_h est la périodicité de la racine caractéristique h.

BIBLIOGRAPHIE

Bergstrom, A.R., (1984), "Monetary, Fiscal and Exchange Rate Policy in a Continuous-Time Econometric Model of the United Kingdom" in *Contemporary Macroecinimic Modelling*, P. Malgrange and P. Muet (editors), London: Basil Blackwell.

sensible aux coefficients qui affectent les variables décalées RS, DNPE et KAP qu'à ceux qui affectent les variables courantes.

Compte tenu que la racine 12 est rattachée plus particulièrement à DNPE et les racines 15 et 16, à DNNE et DNTFPE, il serait intéressant de voir ce que deviennent la stabilité et les propriétés cycliques de la version 3 lorsque ces variables sont transformées en variables exogènes. DNNE et DNTFPE deviennent exogènes lorsque les équations et les coefficients qui s'y rapportent sont exclus de la version 3; on obtient ainsi la version 4. Des données sur certaines racines de la version 4 figurent dans le tableau 1. On retrouve la même stabilité que pour la version 3, si l'on en juge par la racine ayant la valeur la plus élevée. Les cycles de deux et neuf ans sont conservés. En revanche, les cycles de seize et vingt ans, qui, selon l'analyse de sensibilité, étaient étroitement liés aux équations de DNNE, DNPE et DNTFPE, sont disparus. Cela confirme les hypothèses soulevées par l'analyse de sensibilité et montre que si l'on combine une analyse de sensibilité des racines caractéristiques avec une analyse sérieuse des hypothèses économiques sur lesquelles repose le modèle, on peut arriver à déterminer par inférence quels éléments de la structure d'un modèle expliquent les propriétés cycliques de ce modèle.

La fonction de gain pour les principales racines cycliques de la version 4 (racines 3, 14 et 29 selon le tableau 1) est représentée par la courbe en pointillés dans la figure 2. Cette fonction atteint son amplitude maximum à une périodicité de 11 ans et affiche une amplitude presque nulle à une périodicité de 3.7 ans ou moins. Toutefois, contrairement à la fonction de gain de la version 3, elle a une certaine amplitude (faible néanmoins) à des périodicités voisines de 2.

7. DERNIÈRES OBSERVATIONS

Du point de vue méthodologique, les observations tirées des méthodes de simulation classiques sur la stabilité et les propriétés cycliques d'un modèle paraissent simplistes par rapport à celles faites à partir de la méthode analytique. Cette méthode, fondée sur les valeurs propres du modèle linéarisé, a révélé l'instabilité de la version 2 pour des cycles de 14 à 23 ans. De même, elle a permis de constater que la version 3 était instable et qu'elle présentait des cycles de 16 à 22 ans de même qu'un cycle de 2 ans qui était beaucoup plus faible que pouvait le laisser croire le panel 1 de la figure 1. Le cycle de 2 ans observé dans les données peut être considéré dans la structure cyclique du modèle à la condition que les coefficients d'autocorrélation des résidus estimés soient inclus dans la matrice A (matrice des propriétés dynamiques). Tandis que ce cycle de deux annees peut être important pour la prévision, son existence ne sera considérée que dans la mesure où on définira explicitement d'autres variables endogènes et exogènes décalées. Soulignons que si l'analyse des racines a révélé l'existence du cycle de deux ans, ce n'est que par le calcul de la fonction de gain du modèle que l'on a pu constater la faiblesse de ce cycle par rapport aux cycles à longue périodicité.

L'analyse de sensibilité de la période d'une racine nous a montré jusqu'à quel point la méthode analytique pouvait être utile pour déterminer les coefficients de la structure d'un modèle qui sont le plus étroitement liés à l'apparition de cycles particuliers dans le modèle. Tandis que cette analyse a permis d'identifier les coefficients rattachés au stock de capital et au taux d'intérêt à court terme comme d'importants déterminants du cycle de 2 ans, ces mêmes coefficients figurent dans les équations du modèle parce que l'on a tenu compte de l'autocorrélation des résidus. Compte tenu de l'interaction des deux variables précitées, définie par les coefficients structuraux respectifs, on peut difficilement expliquer le cycle de deux ans du point de vue économique. Peut-être que l'autocorrélation des résidus qui a révélé l'existence de ce cycle est un indice de la faiblesse des hypothèses économiques sur lesquelles repose le modèle.

NOTES EXPLICATIVES

1. Les variables stochastiques sont la consommation totale, le taux de croissance de l'emploi à long terme, les prévisions de prix (représentées par le taux de croissance de P à long terme), le taux de croissance de la productivité globale des facteurs à long terme, le niveau des exemptions personnelles en relation avec l'impôt sur le revenu des particuliers, les obligations d'administrations publiques détenues par des étrangers, les dépenses publiques courantes (hormis salaires et traitements), les transferts aux particuliers, l'offre de monnaie centrale, le stock de capital du secteur privé et du secteur public, les stocks du secteur privé, les importations de biens et services, l'emploi dans le secteur privé, le niveau global des prix, le taux de change, la prime de risque pour le capital réel, la prime de risque pour les obligations à long terme, le taux d'intérêt à long terme, le taux d'intérêt à court terme, le taux moyen de l'impôt sur le revenu des sociétés, le taux moyen de l'impôt sur le revenu des particuliers, le taux de rémunération dans le secteur privé et les exportations de biens et services.

2. Par les simulations qu'ils ont effectuées à l'aide du modèle de Brookings, Fromm et coll. (1972) ont observé que le fait de transformer des équations pour tenir compte des erreurs autocorrélées donnait des résultats peu satisfaisants dans les simulations de systèmes complets. Selon eux, l'équation transformée implique un plus grand nombre de valeurs décalées pour les variables endogènes, ce qui fait que dans une simulation dynamique, les valeurs simulées s'écarteront des valeurs réelles par l'accumulation d'erreurs. En ce qui concerne le modèle OBE (Green et coll., 1972) et le modèle de Wharton (Evans et coll., 1972), les transformations autorégressives d'équations ont accru modérément le pouvoir de prévision de ces modèles par rapport à celui des simulations de

divers cycles dans le modèle. Cette fonction sert à exprimer le carré de l'amplitude des cycles observés dans les données relatives aux variables endogènes du modèle en fonction de la périodicité. Puisque seules les amplitudes relatives présentent un intérêt, on normalise les fonctions de gain de manière à ce que l'ordonnée la plus élevée soit égale à 1.

La fonction de gain pour les principales racines cycliques (3, 12, 15, 16, 18, 25 et 34) de la Version 3 (selon le tableau 1) est représentée par la courbe en tirets longs dans la figure 2. L'amplitude maximum correspond à une périodicité de 22 ans. Lorsque la périodicité passe à 11 ans, l'amplitude tombe presqu'à zéro. Mais ce qui est plus important encore, c'est qu'elle est encore tout près de zéro lorsque la périodicité est de 2 ans. La forme de la courbe laisse supposer que le modèle est surtout caractérisé par des cycles à longue périodicité (s'apparentant à une tendance). Alors que les données relatives aux racines caractéristiques de la Version 3 supposent l'existence d'un cycle de deux ans (avec une norme élevée), la fonction de gain, qui situe ce cycle dans l'ensemble du modèle, montre qu'il est négligeable.

Par ailleurs, nous avons pensé qu'il serait intéressant de voir ce qu'est la fonction de gain pour la racine caractéristique qui présente un cycle de deux ans dans la version 3 (racine n^{o} 3); cette fonction est représentée par la courbe en tirets courts dans la figure 2. Nous constatons que l'amplitude de cette fonction est presque nulle à une périodicité de 22 ans et qu'elle atteint une valeur maximum à une périodicité de deux ans. La fonction de gain en soi confirme l'existence d'un cycle de deux ans dans le modèle, mais considéré dans l'ensemble de la structure cyclique du modèle, ce cycle est négligeable.

La figure 2 nous permet aussi de faire une autre constatation; en effet, le périodogramme moyen pour les variables endogènes stochastiques (représenté par le trait continu) ressemble un peu à la fonction de gain qui se rapporte aux racines complexes de la version 3 du modèle. A l'exception des faibles pics enregistrés à la périodicité 3.7 et 2.4, le périodogramme suit un mouvement décroissant de gauche à droite comme la fonction de gain, sauf que la pente est moins accentuée.

6. SENSIBILITÉ DES RACINES CARACTÉRISTIQUES

Dans cette section, nous allons concentrer notre attention sur les racines de la version 3 qui présentent des cycles significatifs pour des périodicités de deux à vingt ans. Nous cherchons ainsi à déterminer quels éléments de la structure du modèle influencent le plus ces racines. A cette fin, nous allons vérifier dans quelle mesure les racines sont sensibles à de faibles perturbations dans les paramètres structuraux. On peut calculer la sensibilité d'une racine du point de vue de la période ou de la valeur absolue. En termes d'élasticité, la sensibilité[7] d'une période s'exprime comme la variation (en pourcentage) de la périodicité par suite d'une variation de 1 % de la valeur de l'élément i, j de la matrice D ou E. La sensibilité de période calculée en fonction de la matrice D décrit la relation entre la période d'une racine particulière et les paramètres des variables endogènes courantes tandis que celle calculée en fonction de la matrice E décrit la relation entre la période d'une racine et les paramètres des variables endogènes décalées. La sensibilité de la valeur absolue d'une racine peut, elle aussi, être calculée en fonction des matrices D ou E. En termes d'élasticité, la sensibilité de la valeur absolue d'une racine est définie comme la variation (en pourcentage) de la valeur absolue d'une racine par suite d'une variation de 1 % de la valeur de l'élément i, j de la matrice D ou E.

Le tableau 2 contient des valeurs de sensibilité pour la période de certaines racines de la version 3. Les lignes du tableau représentent des variables qui figurent dans une équation particulière du modèle. Les deux premières colonnes se rapportent à la variable endogène courante qui figure dans le membre de gauche d'une équation particulière. Les autres colonnes ont trait aux variables endogènes décalées qui font partie des déterminants du membre de droite de cette équation ou d'autres équations. Si nous prenons par exemple la première ligne du tableau, la variable endogène courante, KAP, est la variable du membre de gauche de l'équation de KAP tandis que les variables décalées KAP (-2) et KAP (-3) sont deux des déterminants du membre de droite de la même équation. Selon les données du tableau 2, la racine 3, à laquelle correspond le cycle de deux ans, est liée étroitement au stock de capital (KAP) et au taux d'intérêt nominal à court terme (RS). Une hausse de 1 % du coefficient structurel rattaché à RS entraîne un accroissement de la périodicité de la racine 3 de l'ordre de 3.87 %. De même, une hausse de 1 % du coefficient structurel rattaché à RS (-1) dans l'équation de RS entraînera une diminution de 9.39 % de la période de la racine.

La racine 12, à laquelle correspond le cycle de vingt ans, est surtout liée au taux d'inflation prévu (DNPE) et aux valeurs décalées de cette variable.

Les racines 15 et 16, auxquelles correspond le cycle de 16 ans, se rattachent surtout au taux de croissance de l'emploi à long terme (DNNE) et au taux de croissance de la productivité globale des facteurs à long terme (DNTFPE). Les valeurs élevées indiquent que la période de ces racines est très sensible aux coefficients structuraux rattachés aux variables courantes et décalées DNNE et DNTFPE dans les équations de ces variables.

La racine 18, à laquelle correspond le cycle de 9 ans, est liée aux variables RS, DNPE, KAP et NIC (emploi dans le secteur privé). Elle est cependant plus fortement liée à la variable DNPE. La période de la racine est plus estimés de façon plus efficiente, lorsque regroupés dans la matrice A ils ne produisent pas pour autant une version du modèle qui tienne compte des cycles courts observés dans les données.

Pour démontrer que les cycles courts se retrouvent dans les résidus (e_g) des équations estimées, nous avons calculé ces résidus pour les formes normalisées de (5), qui servent habituellement à la résolution du modèle dans une simulation où les coefficients polynomiaux régressifs sont définis comme nuls. Soulignons que dans beaucoup de cas, ces équations normalisées sont des transformations des équations estimées. Pour des considérations relevant de la théorie économique, nous avons imposé des contraintes aux coefficients estimés de certaines équations. En conséquence, la variable dépendante qui fait l'objet de l'estimation est composée de la variable dépendante normalisée et d'une ou plusieurs variables indépendantes. Par exemple, on peut avoir une équation estimée de forme logarithmique et pour obtenir l'équation normalisée correspondante, il faudra transformer cette équation estimée à l'aide de l'antilogarithme. En règle générale, les résidus des équations normalisées peuvent donc avoir des propriétés très différentes de celles des résidus de l'équation estimée.

L'histogramme du panel 2 de la figure 1 donne la distribution, selon la longueur du cycle, des sommets statistiquement significatifs qui figurent dans les périodogrammes des résidus des équations normalisées pour les variables endogènes stochastiques. Avec un seuil de signification de 0.1, l'hypothèse nulle, selon laquelle les ordonnées de ces périodogrammes sont le résultat d'un bruit blanc, a été rejetée dans 40 cas. Parmi ces 40 cas, 19 ont trait à des cycles de 11 à 22 ans, 6 à des cycles de 2 à 3 ans et 10 à des cycles qui durent de 5 à 11 ans. Si nous comparons les données de cet histogramme aux données du tableau 1 relatives à la version 2, nous voyons que la version AR=0 du modèle ne tient pas plus compte des cycles courts en ce qui concerne les résidus; de fait, on retrouve ce genre de cycles dans les résidus normalisés des équations stochastiques du modèle.

Tandis que la version 2 du modèle utilise des estimations de paramètres plus efficaces, obtenues à l'aide d'un estimateur par les moindres carrés généralisés, les transformations engendrées par (5) pour les variables endogènes et exogènes ne sont pas reflétées dans les équations normalisées qui servent à la résolution du modèle en simulation et, par conséquent, ne sont pas plus reflétées dans les résidue calculés à l'aide de ces équations et utilisés pour construire l'histogramme du panel 2 de la figure 1. Pour pouvoir tenir compte de ces transformations, il faut ajouter des décalages pour les variables endogènes et exogènes dans les équations normalisées du modèle. En conséquence, l'effet des coefficients d'autocorrélation estimés pour les processus d'erreur ne se reflètera que dans la matrice A pour ce qui a trait à la version AR≠0.

Les données du tableau 1 relatives à la Version 3 montrent que la valeur absolue maximum est plus élevée que celle enregistrée pour la version 2 (1.07 contre 1.02). Chose plus importante encore, la racine complexe ayant la norme la plus élevée présente un cycle dont l'amplitude augmente progressivement et dont la périodicité est légèrement supérieure à deux ans. Les racines 12, 15, 16 et 25, qui ont toutes une norme supérieure à 0.87, présentent des cycles stables dont la périodicité varie de 16 à 22 ans. La racine 18, qui a une norme de 0.95, présente un cycle qui a une périodicité de 8.5 ans. Ces observations nous amènent à conclure que la structure cyclique de la version 3 se rapproche plus de celle des données du panel 1 de la figure 1.

Comme la structure cyclique des données relatives aux variables endogènes stochastiques se reflète en majeure partie dans la structure cyclique de la version 3, elle devrait donc être très peu présente dans les résidus des équations normalisées qui servent à la résolution du modèle en simulation. L'histogramme du panel 3 de la figure 1 illustre la distribution des sommets statistiquement significatifs des périodogrammes relatifs à ces résidus. Avec un seuil de signification de 0.1, l'hypothèse nulle, selon laquelle les ordonnées de ces périodogrammes sont le résultat d'un bruit blanc, n'a été rejetée que dans 5 cas. Trois de ces cas représentent des cycles de 22 ans, le quatrième, un cycle de 5 ans et le cinquième, un cycle de 7 ans. L'histogramme n'indique aucun cycle dont la périodicité varie de 2 à 3 ans.

Le comportement cyclique du modèle, tel que le décrivent les racines caractéristiques, ressemble plus à la structure cyclique des données relatives aux variables endogènes stochastiques lorsque les coefficients des modèles autorégressifs des processus d'erreur figurent dans la matrice A.

5. FONCTION DE GAIN, VERSION 3

Dans cette section, nous analysons l'importance des cycles de 2 à 3 ans par rapport aux cycles d'autres périodicités dans la structure cyclique générale de la Version 3. Prise individuellement, la valeur absolue d'une racine en dit peu sur le rôle que joue un cycle particulier dans la structure cyclique générale d'un modèle. Plus la valeur absolue d'une racine sera élevée, plus le cycle correspondant sera ferme. Cependant, lorsque nous sommes en présence d'un modèle à plusieurs racines, la valeur absolue des racines ne permet pas d'évaluer le degré de fermeté d'un cycle particulier parce que la différence de périodicité entre les racines et le nombre de racines qui se situent autour d'une périodicité donnée ont tous deux une influence sur l'amplitude des cycles d'un modèle.

On se sert de la fonction de gain [6] pour évaluer l'importance relative des cycles de deux à trois ans dans la structure cyclique générale de la version 3. La fonction de gain intègre et transforme les données relatives à toutes les racines cycliques de manière à exposer clairement, par des amplitudes relatives, l'importance des

4. PROPRIÉTÉS DYNAMIQUES FONDÉES SUR LES RACINES CARACTÉRISTIQUES

Pour voir comment l'existence d'une autocorrélation entre les résidus influence sur la stabilité et les propriétés cycliques du modèle défini ci-dessus, nous analysons trois versions de ce modèle. Dans le premier cas (version 1), l'estimation se fait par les moindres carrés ordinaires et l'autocorrélation entre les résidus estimés n'est pas prise en considération. Dans le second cas (version 2), l'estimation se fait par les moindres carrés généralisés. On suppose que les résidus suivent un processus autorégressif du second ordre. Un tel processus est nécessaire pour reproduire les cycles observés dans les données relatives aux variables stochastiques du membre de gauche de l'équation. Dans le cas de la version 2, nous nous servons des estimations les plus efficaces des coefficients structuraux pour construire la matrice A. Autrement dit, la version 2 tient compte des effets de la modélisation des erreurs sur les valeurs estimées des paramètres structuraux mais n'intègre pas les coefficients proprement dits de la structure d'erreurs dans la matrice A. Cette version est parfois appelée "version AR=0". La version 3 est identique à la version 2 sauf que les coefficients de la structure d'erreurs sont intégrés dans la matrice A à l'aide de l'équation (5). Cette version est parfois appelée "Version AR≠0".[4]

Le tableau 1 renferme des données relatives aux racines caractéristiques pour les trois versions du modèle. On jugera de la valeur de ces renseignements par la mesure dans laquelle ils sont représentatifs des cycles observés dans les données relatives aux variables endogènes stochastiques du modèle. Ces cycles sont résumés dans l'histogramme du panel 1 de la figure 1.

Cet histogramme donne la distribution, selon la longueur du cycle, des sommets statistiquement significatifs qui figurent dans les périodogrammes des variables endogènes stochastiques.[5] Pour vérifier si l'ordonnée qui correspond à une fréquence donnée dans le périodogramme d'une variable est statistiquement différente de celle qui correspond à une autre fréquence, on utilise la statistique g de Fisher. (Voir Priestley (1981), p. 406-412.) Avec un seuil de signification de 0.1, l'hypothèse nulle, selon laquelle les ordonnées sont le résultat d'un bruit blanc, a été rejetée dans 41 cas. Parmi ceux-ci, 15 représentent des cycles de 11 à 22 ans, 12 représentent des cycles de 2 à 3 ans et les 14 autres représentent, dans des proportions relativement égales, des cycles dont la durée varie de 3 à 11 ans.

Une autre façon de décrire le comportement cyclique des données relatives aux variables endogènes stochastiques est de calculer le périodogramme moyen de ces variables. Ce périodogramme, qui est représenté par le trait continu dans la figure 2, est formé des valeurs moyennes des ordonnées des périodogrammes des diverses variables. Avant de calculer ces valeurs moyennes, on normalise les ordonnées de chaque périodogramme en les divisant par la variance de la série dont on a éliminé la tendance temporelle. Cette opération est effectuée pour deux raisons : premièrement, assurer une meilleure comparabilité des périodogrammes entre les variables et deuxièmement, faire en sorte que l'on puisse comparer le périodogramme moyen à l'histogramme des sommets statistiquement significatifs représenté dans le panel 1 de la figure 1. Après le calcul des valeurs moyennes, on normalise de nouveau les ordonnées en les divisant par l'ordonnée la plus élevée, et cela dans le but de pouvoir comparer le périodogramme moyen aux fonctions de gain analysées dans la section 5 pour diverses versions du modèle (voir aussi la note explicative 6).

Si nous comparons le périodogramme moyen à l'histogramme du panel 1 de la figure 1, nous constatons que la direction du changement, qui indique l'importance des cycles à diverses périodicités, est la même dans huit cas sur dix. Les changements sont de direction contraire uniquement lorsqu'on passe de 7.3 à 5.5 ans et de 3.1 à 2.7 ans. En outre, les ordonnées du périodogramme moyen donnent à croire que les cycles courts sont relativement moins importants que les cycles longs... aspect qui nous échappe lorsque nous examinons uniquement l'histogramme.

Plusieurs constatations ressortent de la comparaison entre l'histogramme du panel 1 de la figure 1 et les données du tableau 1 relatives à la version 1 du modèle. La norme de racine complexe la plus élevée est 1.3 et elle s'accompagne d'un cycle dont l'amplitude croît progressivement et qui a une périodicité de 17.2 ans. Les racines 10, 11 et 12 sont toutes inférieures à un, ce qui suppose que les cycles correspondants, dont la périodicité varie de 14 à 16 ans, sont amortis. Ces cycles correspondent aux cycles de 11 à 22 ans représentés dans le panel 1 de la figure 1. Mais ce qu'il faut surtout retenir de cette comparaison, c'est que, selon les données du tableau 1, la version 1 du modèle n'inclut pas les cycles de 2 à 3 ans observés dans les données relativesauxvariablesendogènesstochastiques.Commecescyclesseretrouventdanscesdonnéesmaisnondanslesracines caractéristiques de la matrice A, qui est calculée à partir des coefficients estimés de la Version 1, on doit pouvoir les observer dans les résiduels estimés du modèle.

Comme nous l'avons indiqué plus haut, la version 2 du modèle permet de considérer l'autocorrélation entre les résidus des équations estimées. Selon le tableau 1, la valeur absolue de la racine réelle la plus élevée passe de 1.13 à 1.02 lorsqu'on change de version; néanmoins, cette racine réelle demeure instable. Quant aux racines complexes, la norme la plus élevée correspond à un cycle de 23.4 années. Les racines complexes qui ont une norme inférieure à un supposent des cycles amortis dont la périodicité est supérieure à 14 ans. Bien que la racine 15 révèle l'existence d'un cycle de trois ans, sa norme peu élevée suppose que l'amplitude de ce cycle est faible par rapport à celle des autres cycles du modèle. Il convient de souligner qu'aucune autre racine complexe dans la Version 2 n'est associée à un cycle court. Nous pouvons en déduire que même si les coefficients de la version AR=0 sont estimés de

$e_{g,t}$ est un des G résidus,

m_g est le décalage maximum en y_g qui figure dans n'importe quelle des G relations fonctionnelles,

n_k est le décalage maximum en x_k qui figure dans n'importe quelle des G relations fonctionnelles.

En supposant que les résidus ont une valeur nulle (valeur espérée), nous pouvons transformer le modèle déterministe de forme structurelle de manière à obtenir l'équation (2), qui est la représentation linéaire de l'espace d'états du modèle,

$$D_t.\Delta z_t = E_t.\Delta z_{t-1} + F_t.\Delta x_t + H_t.\Delta\beta \qquad (2)$$

où z_t, Δx_t et $\Delta\beta$ représentent la différence entre les valeurs des variables et des paramètres d'espace d'états et les valeurs simulées. D_t, E_t, F_t et H_t sont les matrices des dérivées de la représentation d'espace d'états du modèle de forme structurelle par rapport aux variables endogènes de la période courante, aux variables endogènes décalées, aux variables exogènes et aux paramètres respectivement.

Si nous supposons que D est inversible, la représentation linéaire d'espace d'états du modèle de forme réduite est exprimée par l'équation :

$$\Delta z_t = A_t.\Delta z_{t-1} + B_t.\Delta x_t + C_t.\Delta\beta \qquad (3)$$

Comme le laissent voir clairement les équations (2) et (3), LIMO ne s'intéressera qu'aux systèmes d'équations déterministes pour lesquels les résidus des équations stochastiques sont supposés nuls. Il faut donc prévoir une autre méthode pour les systèmes d'équations avec perturbations autocorrélées ou bien il faut transformer ces équations de manière à les rendre compatibles avec le système LIMO.

Supposons que les résidus e_g d'une des variables endogènes stochastiques (y_g) de (1) sont autocorrélés :

$$y_g = f_g + e_g \qquad (4)$$

Si u_t est une variable aléatoire indépendante et P(L), un polynôme en L (opérateur de décalage), nous pouvons exprimer l'autocorrélation entre les e_g par équation:

$$e_{g,t} = p_1 e_{g,t-1} + p_2 e_{g,t-2} + \ldots + u_t \text{ où } e_{g,t} = \frac{u_t}{(1 - P(L))}$$

Si nous multiplions chaque terme de l'équation (4) par le polynôme $(1-P(L))$ et que nous normalisons en fonction de y_g, nous obtenons l'équation:

$$y_g = P(L)y_g + (1 - P(L))f_g + u_t \qquad (5)$$

Il est facile de voir ce qu'implique l'équation (5) pour l'analyse d'un modèle à l'aide du programme LIMO, telle qu'elle est illustrée dans les équations (1) à (3). Il faudra ajouter des décalages pour la variable endogène dans l'équation (1), élargir le vecteur des coefficients β de manière à y inclure les coefficients du polynôme $(1-P(L))$ et ajouter des décalages pour les variables exogènes dans f_g. Lorsque des équations avec erreurs autocorrélées sont envisagées de cette façon, elles peuvent être intégrées à un modèle et servir à l'analyse de sa stabilité et de sa structure cyclique[2] par le programme LIMO.

La matrice A de l'équation (3) est appelée matrice des propriétés dynamiques et elle est le seul lien entre le comportement du modèle linéarisé dans le passé et son comportement dans la période courante. Tandis que ce comportement dépend de la version linéarisée du modèle, Kuh et coll. (1985, p. 14) soulignent les avantages qu'il y a à utiliser la méthode analytique exposée ci-dessus plutôt que d'autres méthodes de simulation non linéaires. Le comportement du modèle est décrit par les racines caractéristiques de A. La valeur absolue des racines indique la stabilité du modèle et la composante imaginaire renferme des renseignements sur ses propriétés cycliques. Une valeur absolue supérieure à un signifie que le modèle est instable, c'est-à-dire qu'il ne retrouve pas son état initial après avoir subi une perturbation. En revanche, un modèle pour lequel la valeur absolue la plus élevée est inférieure à un retrouvera son état initial mais cela se fera d'autant plus lentement que la norme de la racine, même inférieure à un, sera élevée. La composante imaginaire de la racine est utile pour décrire les cycles du modèle. L'amplitude de ces cycles sera croissante, constante ou décroissante selon que la norme de la racine complexe sera supérieure, égale ou inférieure à un. On mesure les cycles par la périodicité, qui est définie comme la longueur d'un cycle complet exprimée dans l'unité de temps (mois, trimestre, année) propre au modèle.[3]

avaient généralement une faible amplitude et étaient fortement amortis. Cette caractérisation des cycles a été rendue possible non seulement grâce à l'observation du comportement des multiplicateurs, mais aussi grâce à l'estimation des spectres des variables endogènes simulées sur une longue période. Cet exercice a été réalisé pour plusieurs variables principales des modèles de Green, Evans et Fromm et pour plusieurs valeurs de perturbations aléatoires autocorrélées.

Dans cet article, nous allons nous servir d'un modèle économétrique représentatif pour démontrer la supériorité des méthodes analytiques par rapport aux méthodes de simulation. Les méthodes analytiques reposent sur le calcul des racines caractéristiques du modèle et la sensibilité de ces racines aux paramètres du modèle. Elles permettent de définir directement la stabilité et les propriétés cycliques du modèle. Du point de vue méthodologique, elles représentent un outil puissant pour l'analyse des données dans le temps et pour les modèles construits à l'aide de ces données, non seulement en économie mais dans beaucoup d'autres branches des sciences sociales.

2. MODÈLE

Le modèle utilisé pour cette analyse est inspiré d'un modèle de simulation mis au point par Duguay et Rabeau (1987, 1988). Comme ceux-ci décrivent leur modèle en détail, nous n'allons mentionner ici que les modifications qui y ont été apportées. Le modèle construit par Duguay et Rabeau représente une économie fermée. Point plus important encore, les valeurs des coefficients sont imposées et non estimées. Elles sont choisies de manière à rendre une image empirique de la structure de l'économie. Le modèle, quant à lui, est structuré pour que l'on puisse modifier facilement les coefficients selon que l'on veut passer d'une perspective Keynesienne à court terme à une perspective néoclassique à long terme (croissance à taux constant) ou vice-versa. Il comprend cinq grandes composantes : demande des particuliers, production et emploi, prix et salaires, secteur des administrations publiques et secteur monétaire et financier.

Le modèle que nous allons utiliser se distingue de celui de Duguay et Rabeau à deux points de vue. Premièrement, il représente une économie ouverte puisqu'il comprend des équations pour les importations et les exportations de biens et services et le taux de change. Les importations et les exportations sont des fonctions du revenu intérieur et du revenu étranger ainsi que des prix relatifs redressés en fonction du taux de change. La variation du taux de change est une fonction de la variation du rapport de la balance commerciale au revenu réel, de la variation de l'écart entre les taux d'intérêt au pays et ceux à l'étranger et de la variation de la portion de la dette publique détenue par des étrangers. Cette formulation, qui s'écarte nettement du modèle de Duguay et Rabeau, ouvre la voie à une autre opération, qui est la vente d'obligations à des étrangers. Cela pourrait être le moyen pour une administration publique de compenser l'effet d'une augmentation de l'intérêt réel. Les obligations d'administrations publiques détenues par des étrangers sont une fonction de cet intérêt réel.

Ce qui distingue aussi notre modèle de celui de Duguay et Rabeau est que les valeurs des coefficients sont estimées plutot que d'être imposées. Cette estimation repose en majeure partie sur des données trimestrielles désaisonnalisées des Comptes nationaux pour la période de 1966 à 1988. Cette méthode permet l'estimation des coefficients et le calcul des résidus de régression.

Le modèle comprend 58 variables endogènes; 35 d'entre elles sont des identités et les 23 autres sont estimées par des méthodes de régression.[1]

3. MÉTHODOLOGIE

Le but de cette étude est d'appliquer l'analyse de modèle linéaire (Linear Model Analysis (LIMO)) au modèle décrit ci-dessus afin d'en illustrer la stabilité et les propriétés cycliques et de montrer ce que deviennent ces caractéristiques lorsque les estimations de paramamètres sont corrigées pour tenir compte de l'autocorrélation des résidus des équations stochastiques. Nous exposons brièvement ici la théorie qui est à la base de LIMO.

Nous pouvons définir un modèle stochastique général de forme structurelle par la formule

$$f(\{y_{g,t}, \ldots, y_{g,t-m_g}\},\{x_{k,t}, \ldots, x_{k,t-n_k}\},\{\beta_j\}) = (e_{g,t}) \qquad (1)$$

où

g $= 1,\ldots,G,\ k = 1,\ldots,K,\ \text{et}\ j = 1,\ldots,J,$

f est un vecteur formé de G relations fonctionnelles,

$y_{g,t}$ est une des G variables endogènes,

$x_{k,t}$ est une des K variables exogènes,

β_j est un des J coefficients constants,

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

LES RÉSIDUS: LE CHAÎNON MANQUANT DANS L'ANALYSE CYCLIQUE DES MODÈLES ÉCONOMÉTRIQUES ET DE LEURS DONNÉES

L. Sager, P. Lin et T. Petersen[1]

RÉSUMÉ

Depuis une vingtaine d'année, on évalue les modèles macro-économiques estimés en comparant de façon critique les résultats non stochastiques de simulations. Ces simulations sont soigneusement concues dans le but de montrer la stabilité et les propriétés cycliques des modèles. Dans cet article, nous proposons une façon mieux équilibrée d'évaluer les modèles économétriques en utilisant à la fois les méthodes de simulation et des méthodes analytiques fondées sur les valeurs propres du modèle, comme le proposait déjà au début des années 1970 Philip Howrey. L'exemple utilisé ici est une version modifiée d'un modèle de simulation macro-économique construit récemment par Duguay et Rabeau (1987, 1988). La création du progiciel LIMO, qui est une composante du programme TROLL (1983), a grandement facilité l'utilisation des méthodes analytiques dans l'évaluation des modèles économétriques. Dans cet article, nous illustrons l'utilisation du progiciel afin de définir les divers cycles d'un modèle et nous nous servons de la fonction de gain du modèle pour évaluer l'importance relative des cycles. Nous associons ensuite ces cycles aux paramètres structuraux du modèle et à d'autres paramètres qui expriment l'autocorrélation des résidus dans les équations estimées.

1. INTRODUCTION

Dans cet article, nous nous servons d'une méthode analytique fondée sur les valeurs propres d'un modèle économétrique pour évaluer dans quelle mesure les résidus d'équations estimées permettent de définir la stabilité et les propriétés cycliques du modèle. Notre article s'apparente d'une part avec les ouvrages de Green et coll. (1972), Evans et coll. (1972) et Fromm et coll. (1972), qui ont recours à des méthodes de simulation, et d'autre part à l'ouvrage de Howrey (1972), qui utilise des méthodes analytiques. Dans la mesure où notre article traite l'évaluation et la comparaison de modèles économétriques, il s'apparente vaguement aux articles qui ont pour sujet le pouvoir de prédiction des modèles économétriques (Burmeister et Klein (1974)) et aux exercices de comparaison de modèles de de Bever (1979) et O'Reilly (1983). Ces études utilisent pour la plupart des méthodes de simulation.

Lorsqu'ils analysent la stabilité et les propriétés cycliques d'un modèle économétrique, la plupart des analystes, hormis des gens comme Howrey (1972)et Bergstrom (1984), procèdent de façon classique en concentrant leur attention sur les multiplicateurs d'impact et les multiplicateurs dynamiques. On mesure l'effet d'une perturbation sur les variables endogènes d'un modèle en faisant une simulation de ce modèle dans le temps. Pour chaque variable endogène, on calcule un multiplicateur dynamique à des périodes données en déterminant tout d'abord la différence entre la solution après perturbation et la solution de référence, puis en exprimant cette différence en pourcentage de la perturbation dans la variable exogène. Une analyse des multiplicateurs dynamiques dans le temps permet de déterminer si le modèle est stable (ses multiplicateurs diminuent), s'il est caractérisé par des cycles (ses multiplicateurs suivent un cycle quelconque) et quelle est la période de ces cycles.

Plus souvent qu'autrement, les multiplicateurs de perturbation ne suivent pas de cycle ou n'affichent que des cycles courts ou longs ayant souvent de très faibles amplitudes. Les cycles longs sont souvent mal définis car les multiplicateurs ne peuvent être calculés sur une période suffisamment longue. Quoi qu'il en soit, il est rare que les multiplicateurs décrivent bien les cycles, courts ou longs, qui se dégagent des données utilisées dans la construction des modèles.

Comme les multiplicateurs de perturbation ne décrivaient pas réellement les cycles longs et courts qui se dégageaient des données utilisées pour construire les modèles, les analystes ont commencé à s'intéresser davantage à la nature de la perturbation. Même si l'on croyait que les perturbations illustraient bien les résidus des équations estimées du modèle, ces résidus présentaient des propriétés d'autocorrélation et de corrélation croisée. Or, le genre de perturbations appliquées aux modèles reflétaient mal ces propriétés. C'est pourquoi on s'est attaché à produire des perturbations aléatoires autocorrélées qui reproduisaient les propriétés statistiques des résidus des équations estimées du modèle pour la période-échantillon. Même si les multiplicateurs issus de l'utilisation de perturbations aléatoires autocorrélées affichaient aussi bien des cycles longs que courts, ceux-ci

[1] L. Sager, P. Lin et T. Petersen, Division des mesures et analyse des industries, Statistique Canada, Ottawa (Ontario), K1A 0T6. Les auteurs remercient D. Rhoades pour ses commentaires éclairants ainsi que V. Campbell pour avoir dactylographié le document.

Baldwin, J.R. et Gorecki, P.K. 1989g. "Dimensions of Labour Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover", document de recherche no 25. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989h. " Measuring Entry and Exit in Canadian Manufacturing Industry: Methodology", document de recherche. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1990. *Structural Change and the Adjustment Process: Perspectives on Firm Growth and Worker Turnover.* Conseil économique du Canada.

Birch, D. 1979. *The Job Generation Process.* Cambridge (Mass): Massachusetts Institute of Technology Program on Neighborhood Change.

Birch, D. 1981. "Who Creates Jobs", *The Public Interest.* no 65, automne, p. 3-14.

Bureau fédéral de la statistique. 1970. *Classification des activités économiques, 1970.* No de cat. 12-501. Ottawa: Approvisionnements et Services Canada.

Johnson, S. et Storey, D. 1985. "Job Generation -- An International Survey: U.S. and Canadian Job Generation Studies Using Dun and Bradstreet Data: Some Methodological Issues", document de recherche no 1. University of Newcastle-upon-Tyne.

McVey, J.S. 1981. *Fusions, ouvertures et fermetures d'usines des grandes entreprises transnationales et autres: 1970-1976.* No de cat. 67-507. Ottawa: Statistique Canada.

Potter, H.D. 1982. "The Census of Manufactures and the Labour Force Survey: some Experimental Approaches to Comparing Establishment and Household Survey Data", un document spécial de la Section de l'analyse et du développement de la Division des industries manufacturières et primaires. Ottawa: Approvisionnements et Services.

Shepherd, W. 1984. "Contestability versus Competition", *American Economic Review* 74: 572-87.

Statistique Canada, 1979. *Concepts et définitions du recensement des manufactures.* No de cat. 31-528. Ottawa: ministère de l'Industrie et du Commerce.

Statistique Canada, 1980. *Industries manufacturières du Canada: niveaux national et provincial, 1978.* No de cat. 31-203. Ottawa: Approvisionnements et Services.

Statistique Canada, 1981. *Industries manufacturières du Canada: niveaux national et provincial, 1979.* No de cat. 31-203. Ottawa: Approvisionnements et Services.

Statistique Canada, 1983. *Organisation des industries et concentration dans le secteur de la fabrication, des mines et de l'abattage, 1980.* No de cat. 31-402. Ottawa: Approvisionnements et Services.

Statistique Canada, 1984. *Industries manufacturières du Canada: niveaux national et provincial, 1982.* No de cat. 31-203. Ottawa: Approvisionnements et Services.

Statistique Canada, 1988. *Building a Longitudinal Database of Firms in the Canadian Economy: The Case of Employment Dynamics.* Ottawa: Approvisionnements et Services.

Storey, D.J. éd. 1985. *Small Firms in Regional Economic Development: Britain, Ireland, and the United States.* Cambridge: Cambridge University Press.

United States Small Business Administration. 1984. "The annual report on small business and competition", tiré du rapport de Ronald Regan, *The State of Small Business: A Report of the President.* Washington (D.C.): U.S. Government Printing Office.

36. Pour plus de détails, consultez le document de recherche de Baldwin et Gorecki (1989h).

37. Dans le cadre de la Classification des activités économiques de 1970, il s'agit de la Division 2, groupe 1, Forêts; de la Division 4, Mines (à l'exception de l'industrie du pétrole brut et du gaz naturel et du groupe 5); et de la Division 5, Industries manufacturières. Pour plus de détails, consultez la publication du Bureau fédéral de la statistique (1970, p. 17). En 1980, la valeur ajoutée des entreprises classées dans le secteur des industries manufacturières se chiffrait à $66,472 millions; à $9,062 millions pour les mines et à $702 millions pour l'exploitation forestière (Statistique Canada, 1983, Tableau explicatif VII, p. 15).

38. La valeur ajoutée du secteur manufacturier permet de classer l'entreprise dans une industrie à quatre chiffres de la CTI en tenant compte de l'entreprise non consolidée la plus importante appartenant à l'entreprise consolidée. Pour plus de détails concernant ces deux concepts d'entreprise, consultez la publication de Statistique Canada (1983, p. 28- 30).

39. En se fondant sur cette définition des entreprises, on dénombrait 30,160 entreprises manufacturières en 1980 (Statistique Canada, 1983, Tableau explicatif VII, p. 15); toutefois, si une entreprise manufacturière correspond exclusivement aux établissements qui la composent et qui sont classés dans le secteur manufacturier, ce secteur compterait alors 30,197 entreprises (Statistique Canada, 1983, Tableau explicatif XIII, p. 21). On dénombrait 37 entreprises minières et forestières ayant des activités dans le secteur manufacturier; par ex., une entreprise minière pouvait posséder une petite fonderie. Par conséquent, en termes de nombre, la définition de l'univers des entreprises manufacturières importe peu.

40. Reportez-vous à la publication de Statistique Canada (1979, p. 43) et au bulletin de J.S. McVey (1981, p. 71).

41. Pour plus de détails, reportez-vous à la publication de Statistique Canada (1983, Tableau explicatif VII, p. 15).

42. Aucun problème semblable ne se pose pour l'entrée et la sortie d'établissements. On suppose qu'un établissement qui quitte le secteur manufacturier (ne remplit pas un questionnaire du recensement annuel des manufactures) n'existe plus. Suivant la terminologie du présent document, il a quitté le secteur par suite d'une fermeture. De la même façon, un établissement ne peut entrer dans le secteur manufacturier que par suite d'une construction.

43. Une autre façon de vérifier si un établissement a rempli un questionnaire du recensement annuel des manufactures pour les années t et t + 1, afin de déterminer si l'entreprise a quitté le secteur en fermant une usine, est de se reporter directement à la question 1.3.2 de ce questionnaire: "Cet établissement a-t-il abandonné les affaires pendant l'année?" La réponse doit nécessairement être "Oui" ou "Non" (Statistique Canada, 1979, p. 79). Selon le groupe de l'Analyse et de l'intégration des micro données d'entreprises, la comparaison des données du questionnaire du recensement annuel des manufactures pour les années t et t + 1 donne un résultat plus fiable que la réponse à la question 1.3.2.

BIBLIOGRAPHIE

Baldwin, J.R. et Gorecki, P.K. 1983. *Entry and Exit to the Canadian Manufacturing Sector: 1970-1979.* Conseil économique du Canada. Document de travail #225, Ottawa, février 1983.

Baldwin, J.R. et Gorecki, P.K. 1987. "Plant Creation Versus Plant Acquisition", *International Journal of Industrial Organisation* 5: 25-41.

Baldwin, J.R. et Gorecki, P.K. 1989a. "Firm Entry and Exit in the Canadian Manufacturing Sector", document de recherche no 23. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989b. "Intra-Industry Mobility in the Canadian Manufacturing Sector", document de recherche. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989c. "Mobility versus Concentration Statistics", document de recherche. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989d. "Productivity Growth and the Competitive Process: the role of firm and plant turnover", document de recherche. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989e. "Mergers Placed in the Context of Firm Turnover", document de recherche. Direction des études analytiques. Statistique Canada.

Baldwin, J.R. et Gorecki, P.K. 1989f. "Job Turnover in Canadian Manufacturing", document de recherche no 22. Direction des études analytiques. Statistique Canada.

13. Les catégories de pays contrôleur correspondaient au Canada, aux E.-U., au R.-U., à d'autres pays européens, et à d'autres pays étrangers.

14. Une entreprise en exploitation est celle qui est classée dans une industrie manufacturière à quatre chiffres pour les années de comparaison initiale et finale.

15. Une usine qui est classée dans la catégorie entrée par acquisition ou sortie par aliénation peut aussi être comprise dans une fusion horizontale. Une telle fusion peut avoir lieu avant ou après l'acquisition.

16. Reportez-vous au document de recherche de Johnson et Storey (1985) pour lire une critique des bases de données de Dun and Bradstreet.

17. Reportez-vous au document de recherche de Storey (1985).

18. Par souci de commodité, le terme "questionnaire abrégé" désigne ici les données des petites usines que l'on a prélevées dans les dossiers administratifs fiscaux au lieu du questionnaire complet postal, ainsi que les questionnaires abrégés proprement dits.

19. Reportez-vous aux publications suivantes de Statistique Canada (1979, p. 44 et 1984, p. xiv). Ces chiffres se rapportent aux "petits" établissements qui semblent remplir pour la plupart des questionnaires abrégés. Consultez la publication de Statistique Canada (1979, p. 43-44).

20. Ces chiffres qui se rapportent aux établissements visés par un questionnaire abrégé pour les années 1970, 1979 et 1982 sont tirés des mêmes sources que celles de la note 19.

21. Consultez la publication de Statistique Canada (1979, p. 42).

22. Selon une étude antérieure de l'entrée/sortie réalisée à l'aide des données du recensement des manufactures qui ne tenaient pas compte des questionnaires abrégés (J.S. McVey, 1981, p. 71).

23. Pour un examen du concept d'activité totale utilisé dans le recensement des manufactures, reportez-vous à la publication de Statistique Canada (1979, p. 21-22). Les mesures des effectifs des entreprises prennent en compte les effectifs totaux, y compris ceux des sièges sociaux, c'est-à-dire, les effectifs des unités auxiliaires ainsi que ceux des établissements d'exploitation.

24. Reportez-vous à la publication de Statistique Canada (1988) où l'on retrouve cette hypothèse.

25. Les mesures qui reposent sur les effectifs totaux se rapprochent beaucoup de celles qui se fondent sur les livraisons.

26. Reportez-vous au document de Baldwin et Gorecki (1983, Tableau 3, p. 15).

27. Reportez-vous au document de Baldwin et Gorecki (1983, Tableau 3, p. 15).

28. Reportez-vous à la publication de Statistique Canada (1979, p. 12-13).

29. Reportez-vous à la publication de Statistique Canada (1979, p. 43-44).

30. L'effort fait par Statistique Canada pour déterminer si un établissement doit être classé dans la catégorie des questionnaires complets varie selon les périodes. Cela aura moins de répercussions sur cette mesure pourvu qu'un établissement qui devient assez important pour recevoir un questionnaire complet, soit éventuellement pris en compte. La possibilité qu'un entrant visé par un questionnaire complet au moment de son entrée dans une industrie ne soit pas saisi par le système à la fin de l'année est plus inquiétante. La qualité des sources de données administratives et la réputation de diligence de Statistique Canada rendent cette éventualité peu vraisemblable.

31. Bien que la ligne de démarcation accuse par la suite une hausse constante, l'augmentation du pourcentage des questionnaires abrégés avant 1983 est relativement faible (4 pour cent seulement). Compte tenu du peu de corrections requises pour les taux d'entrée au moment de la révision de 1975, qui a augmenté de 14 pour cent le nombre d'établissements visés par les questionnaires abrégés, nous n'avons pas poussé plus loin les corrections.

32. Reportez-vous au document de Potter (1982, p. 21).

33. Reportez-vous à la publication de Statistique Canada (1980, p. ix).

34. Reportez-vous à la publication de Statistique Canada (1981, p. x).

35. Nous avons pu constater que les autres hypothèses concernant la distribution des entrants omis n'influaient pas beaucoup sur la moyenne des taux annuels d'entrée et de sortie pour la décennie.

CONCLUSION

Nous avons démontré dans la première section de ce document comment les données longitudinales des entreprises et des usines pouvaient alimenter le débat sur l'importance du processus d'entrée et de sortie. Ce processus ne représente qu'une des forces qui entrent en jeu pour déterminer la vigueur du processus concurrentiel. Pour obtenir un tableau plus complet de la dynamique du processus concurrentiel, il faut aussi vérifier dans quelle mesure la croissance et le déclin des entreprises existantes amènent un changement dans l'importance relative de celles-ci. Il est également essentiel de savoir si les fusions influent beaucoup sur la rotation. Toutes ces questions portent, d'une façon ou d'une autre, sur l'importance et l'origine de la rotation des entreprises. Celles qui portent sur l'effet de la rotation sur la productivité et la rentabilité sont tout aussi importantes.

Pour répondre à ces questions, il faut avoir accès à des bases de données qui peuvent suivre l'évolution des entreprises sur une période donnée. La création de ces bases de données n'est pas facile. Des problèmes compliqués doivent être résolus. Nous nous sommes appliqués à décrire comment, dans le cas de données canadiennes indiquées dans le présent document, nous avons fait face à ces problèmes. Nous espérons que d'autres chercheurs qui se consacrent à la création de bases de données semblables s'y reporteront. Ces problèmes permettront également au lecteur des études qui reposent sur les bases de données dont nous avons parlées, d'évaluer les forces et les faiblesses de la recherche et de comparer, le cas échéant, les résultats de celle-ci avec ceux d'études d'autres pays qui utilisent d'autres sources de données.

NOTES

[1] Une fusion peut, naturellement, entraîner un revirement de situation pour une usine en perte de vitesse et générer ainsi de l'emploi. Des études de la rotation du personnel sur une longue durée pourraient permettre d'examiner dans quelle mesure les usines acquises ont pris de l'expansion comparativement au reste de la population.

[2] Comme tel, il exclut les sièges sociaux et autres unités semblables s'ils sont séparés de l'établissement ou s'ils desservent plus d'un établissement. Pour plus de détails, consultez la publication de Statistique Canada (1979, p. 11- 15).

[3] Un établissement peut se consacrer à plusieurs activités différentes. Pour que l'on classe celui-ci dans le secteur manufacturier, ses activités doivent se concentrer (en se fondant sur la valeur ajoutée) dans ce secteur. Le secteur manufacturier correspond à la Division 5 de la Classification des activités économiques de 1970. Pour plus de détails, consultez la publication du Bureau fédéral de la statistique (1970, p. 23-43).

[4] Dans le cadre du recensement des manufactures, il existe plusieurs types d'unité déclarante, y compris les sièges sociaux et d'autres unités auxiliaires. Nous ne nous intéressons ici qu'aux établissements. Pour plus de détails, consultez la publication de Statistique Canada (1979, p. 10).

[5] Pour plus de détails, reportez-vous aux publications suivantes de Statistique Canada (1979, p. 17-18; 1983, p. 23-25).

[6] Afin de déterminer si une personne juridique en contrôle une autre, nous ne nous intéressons pas seulement aux cas où une société "détient directement ou indirectement plus de 50% des droits de vote de la filiale" (Statistique Canada, 1979, p. 17), mais aussi aux cas de contrôle minoritaire "s'il existe des renseignements concrets à ce sujet ou si une confirmation est obtenue de l'entreprise en question" (Statistique Canada, 1983, p. 25).

[7] Dans certains cas, plusieurs établissements peuvent remplir un questionnaire collectif. Statistique Canada reporte alors les statistiques originales sur chaque établissement qui ont chacun un NSD unique.

[8] Reportez-vous au bulletin de J.S. McVey (1981, p. 72).

[9] Le code d'entreprise logitudinal a été retenu aux fins de l'estimation des statistiques de la concentration des capitaux et de la propriété étrangère, faite par J.S. McVey avec l'aide de J. Bousfield, B. Mersereau et J. Lacroix.

[10] Les résultats indiquent qu'il n'y avait pas beaucoup de différence entre les taux annuels d'entrée et de sortie calculés avec et sans ces critères d'exclusion.

[11] Statistique Canada (1983) et "A Summary of the Establishment Description Tape File", Statistique Canada, document de travail interne non publié, Ottawa, Appendice C-1, p. 2.

[12] Il en va souvent ainsi des bases de données utilisées pour les études américaines et qui sont créées à partir des dossiers de l'assurance-chômage ou de Dun and Bradstreet. Pour un examen des problèmes posés par ces bases de données, reportez-vous à l'ouvrage de Baldwin et Gorecki (1990).

* Choix de l'échantillon

Dans l'analyse qui précède, trois raisons ont été invoquées pour exclure les établissements visés par un questionnaire abrégé. Ces raisons s'appliquent également aux entreprises qui possèdent cette catégorie d'établissements. Ces entreprises exploitent généralement un seul établissement, puisque les établissements appartenant à des entreprises à établissements multiples, qui font partie de plusieurs industries, remplissent toujours un questionnaire complet.[40] Les établissements appartenant à des entreprises à établissements multiples, qui font partie d'une seule industrie, rempliront vraisemblablement aussi un questionnaire complet, puisqu'ils sont importants comparativement à des entreprises à établissement unique.[41] Compte tenu de ces facteurs, nous avons décidé d'exclure les entreprises (a) qui ont toujours possédé un seul établissement (en se fondant sur le code d'établissement multiple/unique) et (b) celles dont l'établissement a toujours rempli un questionnaire abrégé.

L'échantillon est donc composé des entreprises classées dans le secteur manufacturier, à l'exception de celles qui ont toujours possédé un seul établissement qui a toujours rempli un questionnaire abrégé pour le recensement annuel des manufactures.

* Classification des sorties temporaires

Nous avons noté un petit nombre de cas où une usine ou tous les établissements appartenant à une entreprise n'avaient pas fait de déclaration pour une année donnée, mais l'avaient fait pour l'année précédente et suivante. Si nous appliquions à ces cas la ligne de conduite décrite plus haut, nous les aurions classés comme des sortants et, ultérieurement, comme des entrants, et non comme des usines ou des entreprises en exploitation. Des fonctionnaires de Statistique Canada ont proposé plusieurs raisons pour expliquer cette non-déclaration: un incendie, une grève, une révision majeure du matériel ou un ralentissement des activités. Nous avons reclassé ces cas et considéré l'usine ou l'industrie comme étant en exploitation et non comme un sortant ou un entrant.

* Détermination du type d'entrée et de sortie

Nous ne nous sommes pas arrêtés, dans la définition de l'entrée et de la sortie d'entreprises donnée dans la section précédente, à faire la distinction entre les différents types d'entrée et de sortie. Comme nous l'avons déjà décrit, une entreprise peut quitter le secteur manufacturier en fermant toutes ses usines, ou en vendant celles-ci à une autre entreprise (aliénation). De la même façon, une entreprise peut entrer dans le secteur manufacturier en construisant une usine ou en achetant les usines d'autres entreprises (acquisition). Aux fins de l'analyse de l'entrée et de la sortie d'entreprises sur une longue durée, nous avons différencié ces différents types d'entrée et de sortie.[42] La distinction a aussi été faite pour les estimations qui portent sur une courte durée.

Dans le cas des entrées sur une courte durée, nous avons procédé de la façon suivante pour déterminer si une entrée d'entreprise avait eu lieu par suite d'une acquisition, par opposition à une création d'usine: si, d'après le premier questionnaire qu'il a rempli pour le recensement annuel des manufactures, l'entrant possédait un (des) établissement(s) qui existait(ent) déjà l'année précédente, on considérait que l'entreprise était entrée dans l'industrie par suite d'une acquisition; si les établissements n'existaient pas l'année précédente, on classait l'entreprise comme un entrant par suite d'une création. Nous avons procédé de la même façon pour distinguer les types de sortie: si, d'après le dernier questionnaire qu'il a rempli pour le recensement annuel des manufactures, le sortant possédait des usines et que celles-ci existaient toujours l'année suivante (mais avaient changé de propriétaire), on considérait que l'entreprise avait quitté l'industrie du fait d'une aliénation; si, l'année suivante, les usines n'avaient pas rempli le questionnaire, on classait l'entreprise comme un sortant du fait d'une fermeture d'usines.[43]

Un problème peut se poser si une entreprise entre dans l'industrie en acquérant des usines et en en construisant, ou si une entreprise quitte l'industrie en aliénant une usine et en en fermant une. On pourrait résoudre ce problème en comptant l'entreprise deux fois ou en créant une nouvelle catégorie (entrée par suite d'une acquisition et d'une ouverture d'usines). On pourrait aussi classer cette entreprise dans l'une des deux catégories en fonction de l'importance des usines créées comparativement à celles qui sont acquises.

Nous pouvons vérifier les conséquences de l'utilisation de la première méthode à l'aide des données qui ont servi à mesurer l'entrée et la sortie sur une longue durée. Bien que l'entrée de certaines entreprises fût le résultat d'une création et d'une acquisition d'usines, il n'y avait presque pas de chevauchement. Nous avons obtenu ces données en comparant l'état des entreprises en 1970 et en 1979, une période qui couvre une décennie complète. La possibilité pour une entreprise d'entrer dans l'industrie d'une façon puis de prendre de l'expansion de l'autre devrait être plus grande pour une période de dix années que pour la période d'un an qui sert à mesurer l'entrée sur une courte durée. Par conséquent, si l'on fait une étude qui se fonde sur des données annuelles, le chevauchement entre les deux type d'entrée devrait être grandement réduit.

Compte tenu de ce qui précède, nous avons jugé opportun de tenir compte des entrées par suite d'une création et des entrées par suite d'une acquisition d'usines. Par conséquent, nous avons classé les entrants dans l'une ou l'autre des deux catégories d'entrée en fonction des effectifs des usines créées par opposition à ceux des usines acquises. Dans le cas d'une entreprise à établissements multiples, nous avons pris soin de faire le bon choix.

76 et 1980-81. L'augmentation de la taille moyenne des établissements s'est produite brusquement au moment de la reclassification de ceux-ci en 1975 dans la catégorie des questionnaires complets ou dans celle des questionnaires abrégés.

Afin de pouvoir calculer l'effet de la redéfinition de 1975 sur les taux d'entrée estimés, la distribution des entrants de 1973-74 a été tronquée en éliminant les plus petits entrants jusqu'à ce que la taille moyenne de ceux qui restent soit égale à la taille de l'entrant moyen après 1975. Pour ce faire, il a fallu enlever en moyenne 32.1 pour cent des entrants qui représentaient 4.5 pour cent des employés de tous les entrants. Cela correspond à l'estimation de la réduction en pourcentage des chiffres d'entrée antérieurs à 1975 qu'il faut faire pour que ces chiffres soient comparables à ceux du reste de la période.[31]

L'adoption de l'échantillon visé par le questionnaire complet crée un autre problème de mesure. La couverture de l'échantillon d'établissements diminue d'année en année reflétant le recours aux questionnaires complets dans la présente étude ainsi que la baisse graduelle de leur importance en termes de nombre d'établissements. Ce phénomène ne devrait pas avoir trop de répercussions sur le taux d'entrée et de sortie lorsque celui-ci est calculé comme un proportion du nombre d'entreprises ou d'établissements à un moment précis. L'écart est encore moins accentué lorsque le taux d'entrée et de sortie est mesuré en termes d'effectifs, en raison de la taille relativement petite des établissements visés par le questionnaire abrégé. Néanmoins, le calcul des taux annuels d'entrée à l'aide de l'échantillon du questionnaire complet ne porte que sur la période qui se termine en 1982. Après l'an 1982, l'échantillon n'a pas un nombre d'années suffisant pour rendre parfaitement compte de la transition d'un questionnaire abrégé à un questionnaire complet pour un établissement. Par conséquent, les taux d'entrée seront de plus en plus sous-estimés. Ce problème pourra être surmonté lorsque nous disposerons de plus de données.

* Variation de la couverture du recensement

Le second problème a surgi en raison d'un changement majeur de la couverture du recensement. Si ce problème n'était pas résolu, le changement aurait entraîné une fausse augmentation de l'entrée pour les années 1978 et 1979 et une sous-estimation de l'entrée et de la sortie au cours des années précédentes.

Un changement majeur a été apporté en 1978 à la couverture du recensement canadien des manufactures. Statistique Canada perdait en 1972 une source de renseignements administratifs qui permettait d'identifier de nouveaux établissements potentiels.[32] Il en est résulté une réduction de la couverture qui n'a pas été corrigée avant 1978 et, dans une moindre mesure, 1979. En 1978, par exemple, 3,820 nouveaux établissements, qui existaient déjà selon Statistique Canada, ont été ajoutés au recensement des manufactures. Ces "nouvelles" unités représentaient 12 pour cent de tous les établissements dénombrés en 1978; toutefois, la plupart étaient très petits, et l'augmentation des livraisons du secteur manufacturier attribuable à cet ajout était peu important (ces "nouveaux" établissements ne représentaient que 1.7 pour cent du total des effectifs en 1978).[33] En 1979, suite aux améliorations apportées à la couverture, on a ajouté 1,142 autres établissements préexistants qui ne représentaient que 3.3 pour cent de tous les établissements et que 0.37 pour cent du total des effectifs.[34] Afin de résoudre le problème occasionné par le changement de couverture, nous avons tenu compte du nombre d'entrants et d'emplois chez ces derniers que l'on pouvait attribuer à l'élargissement de la couverture pour corriger les taux d'entrée et de sortie.

La correction des taux de 1978 et de 1979 était facile à faire. Nous avons soustrait la valeur du chevauchement.

Pour ce qui est des années précédentes, la correction était plus compliquée. En raison du taux élevé de disparitions chez les nouveaux entrants, l'application pure et simple des chiffres de la couverture élargie de 1978 et 1979 aux années précédentes aurait entraîné une sous-estimation des entrées pour ces années. Pour venir à bout de ce problème, nous avons avancé deux hypothèses: premièrement, nous avons supposé que toutes les entrées oubliées étaient réparties, entre 1972 et 1977, proportionnellement à celles effectivement déclarées,[35] et, deuxièment, que le rythme de disparition hâtive des entrants oubliés correspondait à celui qui avait été observé pour les nouveaux entrants. De cette façon, nous pouvions estimer le nombre d'entrants oubliés chaque année entre 1972 et 1977. Pour calculer les effectifs correspondants, nous avons supposé que le nombre d'employés de chaque entrant oublié équivalait à la moyenne obtenue pour les entrants qui avaient vraiment été pris en compte.

Nous avons également révisé les données des taux de sortie pour tenir compte du fait que la couverture réduite de l'entrée au milieu des années 1970 aurait entraîné une baisse graduelle des taux de sortie calculés. Nous avons eu recours encore une fois au taux de sortie des nouveaux entrants pour les appliquer aux autres entrants. Les corrections n'ont pas beaucoup modifié le taux moyen d'entrée et de sortie calculé pour la décennie.[36]

La base de données annuelle des entreprises

Une entreprise se définit comme tous les établissements des industries manufacturières et primaires[37] sous contrôle commun. Si l'activité d'une entreprise[38] se concentre davantage dans une industrie manufacturière à quatre chiffres que dans toute autre industrie à quatre chiffres du secteur des mines et de l'exploitation forestière, on classe l'entreprise dans le secteur manufacturier. L'échantillon d'entreprises que nous avons utilisé pour la base de données de courte durée comprend celles qui sont classées dans ce secteur.[39]

(50.3). Nous retrouvons plus ou moins le même chevauchement du côté des entrées lorsque nous examinons le nombre d'entreprises en exploitation en 1979.

Mesure de l'entrée et de la sortie sur une courte durée

Comme nous l'avons indiqué précédemment, deux bases de données ont été créées pour mesurer l'entrée et la sortie sur une courte durée qui se rapportent au secteur manufacturier. La première suit d'année en année l'évolution des établissements au cours de la période 1970-82 et la seconde fait de même pour les entreprises. Les bases de données de courte durée permettent uniquement de mesurer l'entrée et la sortie à un niveau élevé de l'agrégration des industries (l'ensemble du secteur manufacturier).

Plusieurs problèmes se sont posés au moment de mesurer l'entrée et la sortie sur une courte durée. Nous les examinerons dans les deux sections suivantes. La première traite de la base de données des établissements et la seconde, de la base de données des entreprises.

La base de données annuelle des établissements

* Choix de l'échantillon

On peut générer les données de l'entrée et de la sortie en tenant compte de tous les établissements ou uniquement de ceux qui remplissent le questionnaire complet ou le questionnaire abrégé. Nous avons décidé d'utiliser seulement les questionnaires complets parce que, entre autres problèmes, la couverture continuellement changeante des questionnaires abrégés produirait des taux d'entrée et de sortie trompeurs[28] (surtout dans le cas de la mesure des taux annuels d'entrée et de sortie). Nous avons démontré dans une section précédente que les données du questionnaire complet se rapprochaient beaucoup des résultats du recensement total pour la période prolongée de 1970 à 1979 (du moins lorsque la mesure de l'entrée est établie selon le nombre de livraisons ou d'emplois visés).

Dans le cas de la base de données de courte durée, l'utilisation exclusive des questionnaires complets comme critère d'échantillonnage est insuffisante. La ligne de démarcation entre questionnaires abrégés et questionnaires complets a changé au fil des années. Pour cette raison, l'utilisation exclusive des données du questionnaire complet produirait certains changements dans l'entrée et la sortie attribuables uniquement à une reclassification. Nous avons résolu ce problème en prenant comme échantillon longitudinal des établissements tous ceux qui avaient rempli un questionnaire complet au moins une fois. Par conséquent, un établissement est considéré comme un entrant pour une année particulière parce qu'il a fait son entrée dans une industrie cette année-là et recevait déjà un questionnaire complet ou en aura rempli un à une date ultérieure.

Cette technique permet d'atténuer sans toutefois éliminer les problèmes causés par la ligne de démarcation mouvante entre les deux types de questionnaire. Elle contribue essentiellement à aplanir les fluctuations en éliminant la composante la plus volatile, les établissements qui bordent cette ligne. Comme les déplacements de la ligne de démarcation sont généralement petits, cette technique est suffisante la plupart du temps; toutefois, des changements importants ont été apportés à la couverture du recensement à deux occasions. Il a alors fallu corriger les estimations d'entrée et de sortie.

* Révision importante, en 1975, de la couverture visée par le questionnaire complet

La démarcation entre le questionnaire abrégé et le questionnaire complet a été révisée en profondeur en 1975.[29] Au début des années 1970, Statistique Canada a haussé lentement la ligne de démarcation de manière à conserver à peu près le même pourcentage d'établissements dans chaque catégorie. Toutefois, en 1975, l'organisme l'élevait considérablement afin de réduire le fardeau de réponse des fabricants plus petits. En conséquence, le pourcentage des établissements qui remplissaient un questionnaire abrégé est passé de 36.1 pour cent en 1974 à 50.1 pour cent en 1975. Il n'a connu aucune augmentation semblable par la suite, bien qu'il ait augmenté lentement au fil des années. Avant 1983, il représentait 54.9 pour cent de tous les établissements comparativement à 50.1 pour cent en 1975. Le pourcentage des employés travaillant dans les établissements visés par le questionnaire abrégé accusait une légère augmentation (7.6 à 8.7 pour cent) au cours de la même période.

La reclassification de la démarcation entre le questionnaire complet et le questionnaire abrégé en 1975 aura moins de répercussions sur les estimations de l'entrée et de la sortie grâce à l'échantillon du questionnaire complet modifié adopté ici.[30] Cela est dû au fait que les établissements qui faisaient partie de la catégorie des questionnaires abrégés au moment de leur entrée en 1975, mais qui ont ensuite pris de l'expansion pour passer dans la catégorie des questionnaires complets (même si la tâche était plus lourde en 1975 en raison de la hausse de la ligne de démarcation qui servait à définir les établissements visés par les questionnaires complets) seront quand même pris en compte. Toutefois, le problème n'est pas éliminé pour autant. Les établissements qui auraient fait la transition d'un questionnaire abrégé à un questionnaire complet en vertu de la définition antérieure à 1975 de ce dernier, mais qui ne le font pas en vertu de la nouvelle définition, seront omis.

L'augmentation de la taille moyenne des établissements entrants qui s'est manifestée après 1975 témoigne d'une certaine réduction de l'entrée mesurée attribuable à la révision survenue cette année-là. Chacun de ces établissements employaient en moyenne 20 personnes entre 1970-71 et 1972-73, mais 28.1 personnes entre 1975-

Tableau 3

Importance des catégories d'entrée dans les industries manufacturières canadiennes entre 1970 et 1979, établie à partir d'autres ensembles de données: (calcul correspondant à la moyenne obtenue pour 167 industries à quatre chiffres)

Catégorie	Part des usines		Part des livraisons	
	Univers intégral	Echantillon question. complet	Univers intégral	Echantillon question. complet
Toutes	100.0	100.0	100.0	100.0
Toutes les entrées d'entreprises selon				
1) Ouverture d'usine (23)	36.9	18.8	14.4	11.5
2) Acquisition (22)	6.5	8.7	10.4	10.7
3) Transfert d'usines (26)				
a) même propriétaire	3.5	4.7	3.3	3.5
b) nouveau propriétaire	0.6	0.9	1.0	1.1
Toutes les entreprises en exploitation				
4) Etablissements en exploitation (15)	46.8	59.2	63.0	65.0
5) Acquisition d'une usine (12)	1.6	2.2	2.8	3.0
6) Nouvelle usine (13)	3.6	4.6	4.2	4.4
7) Transfert d'usines (16)	0.5	0.7	0.9	0.9

Remarque: 1) Se reporter au Tableau 1 et au texte pour connaître la définition des catégories.
2) On mesure l'importance des différentes catégories d'entrée selon le nombre d'usines ou les livraisons des usines appartenant à des entreprises d'une catégorie particulière, exprimés comme un pourcentage de toutes les usines ou de toutes les livraisons dans une industrie.
3) Le calcul de la moyenne tient compte de toutes les industries (y compris celles dont la valeur dans une catégorie particulière est égale à zéro).

Tableau 4

Importance des catégories de sortie dans les industries manufacturières canadiennesentre 1970 et 1979, établie à partir d'autres ensembles de données (calcul correspondant à la moyenne obtenue pour 167 industries à quatre chiffres)

Catégorie	Part des usines		Part des livraisons	
	Univers intégral	Echantillon question. complet	Univers intégral	Echantillon question. complet
Toutes	100.0	100.0	100.0	100.0
Toutes les sorties d'entreprise selon				
1) Fermeture d'usines (34)	32.4	24.6	14.1	13.3
2) Aliénation (31)	8.5	10.0	12.5	12.7
3) Transfert d'usines (37)				
a) même propriétaire	3.8	4.3	3.4	3.5
b) nouveau propriétaire	0.6	0.8	1.3	1.3
Toutes les entreprises en exploitation				
1) Etablissements en exploitation (15)	50.6	55.3	62.9	63.4
2) Aliénation (11)	0.5	0.6	1.1	1.1
3) Fermeture d'usines (14)	3.3	3.8	3.7	3.8
4) Transfert d'usines (17)	0.4	0.5	0.6	0.8

Remarque: 1) Se reporter au Tableau 1 et au texte pour connaître la définition des catégories.
2) On mesure l'importance d'une catégorie selon le nombre d'usines ou les livraisons des usines dans cette catégorie, exprimés comme un pourcentage de toutes les usines ou de toutes les livraisons.
3) Le calcul de la moyenne tient compte de toutes les industries.

acquisition (catégorie 22). Ce groupe représente environ 10 pour cent de la catégorie de l'entrée par suite d'une acquisition qui ne comprend pas les transferts d'usines (rangée 2, Tableau 3). On range dans la seconde sous-catégorie (rangée 3a, Tableau 3) les usines qui ont gardé le même propriétaire (3.5 pour cent de toutes les livraisons). Ces usines font partie de la catégorie d'entrée de nouvelles entreprises par suite de l'ouverture d'une usine (catégorie 23). Leurs livraisons représentent environ 30 pour cent de la catégorie d'entrée de nouvelles entreprises par suite de la construction d'une usine lorsque les transferts ne sont pas pris en compte (rangée 1, Tableau 3).

Les résultats des sorties sont le reflet de ceux des entrées. Les transferts où n'intervient pas de changement de propriétaire peuvent augmenter d'environ 30 pour cent le taux de sortie des entreprises par suite de la fermeture d'une usine. Le nombre de transferts d'usines par des entreprises en exploitation est également élevé par rapport au nombre d'ouvertures d'usines par ce type d'entreprise. Les transferts représentent 0.9 pour cent des livraisons en 1979 (rangée 7, Tableau 3) comparativement à 4.4 pour cent en moyenne (rangée 6, Tableau 3) pour la part des usines neuves de la catégorie des entreprises en exploitation. Il faut donc tenir compte des transferts puisque ceux-ci peuvent modifier sensiblement les taux d'entrée et de sortie calculés sur une longue durée.

* Chevauchement des catégories d'entrée et de sortie

Afin de pouvoir examiner l'étendue de ce problème, nous avons estimé le nombre d'établissements et le nombre d'entreprises dans différentes catégories d'entrée pour un échantillon réduit de 141 industries (un échantillon qui avait servi à une analyse de régression pour les entrants (Baldwin et Gorecki, 1987)). Nous n'avons tenu compte que des établissements visés par un questionnaire complet.

Pour chaque industrie comprise dans l'échantillon de 141 industries manufacturières à quatre chiffres, le nombre moyen d'entreprises classées entrants était de 24.6 avant 1979, 4.9 entreprises par suite d'une acquisition et 21.7 par suite de la construction d'une usine.[26] Par conséquent, parmi les 24.6 entrants, 2 entreprises en moyenne ont fait leur entrée pendant la période 1970-79 en acquérant une usine et en en construisant une. Pour ce qui est des sorties, en moyenne 38.3 entreprises existantes en 1970 ont cessé leurs activités au cours de la décennie dont 7.2 par suite d'une aliénation et 33.2 suite d'une fermeture. Ainsi, parmi les 38.3 entreprises sortantes, environ 2.1 par industrie en moyenne étaient dues à une aliénation et à une fermeture d'usine au cours de la période 1970-79.[27]

Dans la catégorie des entreprises en exploitation, 50.3 entreprises en moyenne possédaient une usine dans l'industrie au cours des années initiale et finale. On dénombre 49.8 entreprises possédant des usines qui sont demeurées dans l'industrie au cours de la décennie, 1.6 qui ont aliéné des usines et 3.7 qui en ont fermées. La somme des sous-catégories (55.1) est d'environ 10 pour cent supérieur au nombre d'entreprises en exploitation usine, ou en faisant les deux. Les entreprises en exploitation peuvent construire une usine, aliéner une usine ou en acquérir une. Cela peut créer plusieurs difficultés. Les pourcentages des différentes catégories ne totalisent plus 100 pour cent. Les différents degrés d'activité dans plusieurs catégories peuvent alors avoir une influence sur les comparaisons de l'importance de l'entrée pour toutes les industries. Nous avons examiné la portée de ce problème en nous servant de la base de données de longue durée.

PROBLÈMES DE MISE EN APPLICATION: PRÉCISIONS

Dans les sections précédentes, nous avons décrit et examiné d'une manière générale les choix à faire concernant la production de trois bases de données visant à étudier différents aspects de l'entrée et de la sortie, ainsi que les définitions qui s'y rapportent. Nous traiterons plus en détail dans les sections suivantes de chacune de ces bases de données en mettant l'accent sur la façon de résoudre les problèmes de mise en application.

La base de données de longue durée

Sur une courte durée, les composantes cycliques et stochastiques de la croissance et du déclin des entreprises tendent à écraser les tendances structurelles. Il en va de même pour l'entrée et la sortie. Parce que nous avons jugé que l'importance de l'entrée et de la sortie ne pourrait se manifester que sur une plus longue durée, nous avons utilisé la base de données qui offre le plus de détails, soit la base de données de longue durée. L'entrée et la sortie sont mesurées au niveau de l'industrie à quatre chiffres. Toutes les catégories du Tableau 1 sont utilisées.

* Choix de l'échantillon

Comme nous l'avons déjà indiqué, l'importance de l'entrée et de la sortie peut être évaluée à l'aide de l'univers intégral des entreprises et des établissements ou de l'échantillon réduit visé par le questionnaire complet. L'utilisation exclusive de l'échantillon réduit comporte des avantages. Toutefois, avant de recourir largement à cet échantillon, il est essentiel d'évaluer les effets de cette démarche.

Le Tableau 2 indique le pourcentage de toutes les industries à quatre chiffres pour lesquelles on a enregistré au moins une observation dans chacune des catégories d'entrée et de sortie. Les ratios de couverture sont présentés pour l'ensemble des établissements compris dans chaque industrie et pour seulement l'échantillon visé par le questionnaire complet. Le fait de choisir cet échantillon n'influe de toute évidence pas beaucoup sur la couverture.

Le Tableau 3 renferme deux estimations de l'importance des différentes catégories d'entrée établies à partir du nombre d'établissements et de la valeur des livraisons. Le Tableau 4 contient des estimations de l'importance des catégories de sortie qui se fondent sur les deux échantillons. Dans chaque cas, la première estimation repose sur l'univers intégral, et la seconde, sur l'échantillon qui remplit le questionnaire complet. L'importance d'une catégorie d'entrée et de sortie est mesurée en fonction des totaux de l'ensemble utilisé (toutes les observations dans le premier cas, seulement celles de l'échantillon visé par le questionnaire complet dans le second). Les estimations qui figurent dans les Tableaux 3 et 4 représentent la moyenne de l'importance de chaque catégorie pour 167 industries à quatre chiffres.

Il est évident que lorsqu'on tient compte du nombre d'établissements, l'utilisation de l'échantillon visé par le questionnaire complet a une influence sur l'importance de l'entrée et de la sortie; l'effet est beaucoup moindre si l'on tient compte des livraisons. Cet échantillon peut donc servir à mesurer les valeurs des livraisons qui ont été modifiées par l'entrée et la sortie sans trop les fausser. Cette conclusion s'applique également à d'autres mesures d'intrant ou d'extrant.[25]

* Entrée d'usines entièrement nouvelles v. entrée d'usines reclassées

Etant donné que l'on peut définir une entrée et une sortie d'établissement en tenant compte ou non du transfert d'usines d'une industrie à une autre, nous avons examiné l'importance de la catégorie du transfert d'usines. A cette fin, tous les établissements en exploitation auxquels on avait attribué un code de la CAE en 1979 différent de celui de 1970 ont été considérés comme des entrants en 1979 et des sortants en 1970 de l'industrie à quatre chiffres en question de la CAE, selon un transfert d'usines. Nous avons divisé les transferts d'usines en deux catégories: les transferts associés à une entrée d'entreprises et ceux associés à des entreprises en exploitation. Dans le premier cas, le transfert a permis d'intégrer une nouvelle entreprise dans une industrie; dans l'autre cas, l'entreprise, dont l'usine a reçu un nouveau code de la CAE, possédait déjà une usine dans cette industrie. On considère qu'une entreprise est en exploitation ou nouvelle selon qu'elle possède ou non une usine dans une industrie particulière à quatre chiffres.

Les Tableaux 3 et 4 renferment également des estimations de l'importance de l'entrée de nouvelles entreprises par suite d'un transfert. Le taux d'entrée de nouvelles entreprises par suite d'un transfert s'élève à 4.6 pour cent lorsqu'on utilise les livraisons et l'échantillon visé par le questionnaire complet pour mesurer. L'échantillon choisi n'influe pas beaucoup sur ce taux. On peut diviser en deux sous-catégories le taux d'entrée des nouvelles entreprises dans une industrie par suite d'un transfert: la première (rangée 3 b, Tableau 3) correspond aux transferts dans le cas d'un changement de propriétaire (1.1 pour cent de toutes les livraisons de l'industrie selon l'échantillon visé par le questionnaire complet) et sont inclus dans la catégorie de l'entrée par suite d'une

Nous n'avons pas suivi cette ligne de conduite dans le cadre de la présente étude. Nous pensons qu'il y a suffisamment de retard dans les déclarations du recensement pour considérer que les chiffres des effectifs totaux pour la première et la dernière année de déclaration d'un établissement correspondent essentiellement à une année complète d'opération. Nous en avons fait la vérification en examinant les chiffres des effectifs déclarés par les entreprises sortantes, pour l'année de sortie et l'année précédente. Les différences étaient relativement mineures et sûrement pas de l'ordre de 100 pour cent, comme le laisserait supposer la règle du doublement.

- Nuances dans les définitions

Une fois que l'on a déterminé les catégories à mesurer, il reste à résoudre les problèmes de mise en application parce que, dans certains cas, d'autres définitions peuvent servir à mesurer une catégorie d'entrée et de sortie particulière. Deux questions ont été examinées attentivement. Premièrement, faudrait-il considérer les usines qui changent de secteur industriel comme des créations et des disparitions; et, deuxièmement, y a-t-il chevauchement des catégories "entrée d'entreprises" et "sortie d'entreprises"?

Reclassification d'une usine considérée comme une entrée. L'entrée d'un établissement correspond à l'apparition d'une nouvelle usine. Une nouvelle usine peut figurer dans une industrie particulière à quatre chiffres parce qu'elle n'était pas prise en compte dans le recensement des manufactures, ou parce qu'elle faisait partie d'une autre industrie et qu'elle a été transférée dans l'industrie donnée. Un établissement est classé dans une industrie en fonction de ses produits. Lorsqu'il change son type de produit, l'industrie à laquelle il appartient aux fins du recensement peut changer (bien que, pour ce faire, l'on attende de vérifier si le changement de produit est définitif). Un transfert a lieu parce qu'une usine dont la production principale était attribuée auparavant à d'autres industries concentre maintenant la plus grande partie de ses activités sur des produits qui appartiennent à l'industrie en question.

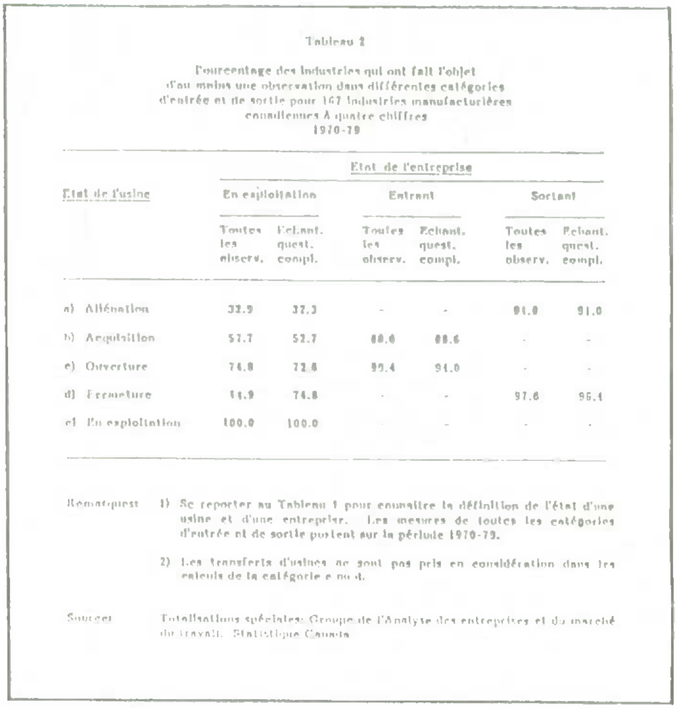


Tableau 2

Pourcentage des industries qui ont fait l'objet d'au moins une observation dans différentes catégories d'entrée et de sortie pour 167 industries manufacturières canadiennes à quatre chiffres
1970-79

Etat de l'usine	Etat de l'entreprise					
	En exploitation		Entrant		Sortant	
	Toutes les observ.	Echant. quest. compl.	Toutes les observ.	Echant. quest. compl.	Toutes les observ.	Echant. quest. compl.
a) Aliénation	33.9	37.3	-	-	91.8	91.0
b) Acquisition	57.7	52.7	88.0	88.6	-	-
c) Ouverture	74.8	72.8	99.4	94.0	-	-
d) Fermeture	44.9	74.8	-	-	97.6	96.4
e) En exploitation	100.0	100.0	-	-	-	-

Remarques: 1) Se reporter au Tableau 1 pour connaître la définition de l'état d'une usine et d'une entreprise. Les mesures de toutes les catégories d'entrée et de sortie portent sur la période 1970-79.

2) Les transferts d'usines ne sont pas pris en considération dans les calculs de la catégorie c ou d.

Source: Totalisations spéciales: Groupe de l'Analyse des entreprises et du marché du travail, Statistique Canada

Il n'est pas facile de déterminer la ligne de conduite à suivre dans le cas des usines qui ont été reclassées dans une autre industrie. La reclassification dans l'industrie N des usines de l'industrie M entraine pour le recensement un transfert des effectifs totaux de M à N qui n'est généralement pas associé à la création de nouveaux emplois dans N équivalant au total de l'emploi de l'usine reclassée. Par conséquent, les mesures de l'entrée qui tiennent compte de cette forme d'entrée dans N semblent, à première vue, surestimer le taux de création et de perte d'emplois attribuable à l'entrée et à la sortie. Ce raisonnement laisserait supposer qu'il faudrait, dans le cas des études de la rotation du personnel, exclure les transferts.

D'autre part, dans le cas des études de la concurrence, les transferts sont importants et devraient être inclus dans ces études parce qu'ils intègrent de nouveaux participants dans l'industrie.

Nous avons résolu le problème en mesurant l'importance des transferts à l'aide de la base de données de longue durée. Pour ce qui est des bases de données de courte durée, les transferts posent probablement moins de problèmes. Les bases de données qui servent à mesurer les taux annuels d'entrée prennent seulement en compte les entrées et les sorties du secteur manufacturier dans son ensemble. Les transferts d'une industrie à quatre chiffres à une autre au sein de ce secteur ne posent pas alors de problèmes. Toutefois, au niveau de l'agrégation, des transferts d'entrée et de sortie peuvent se produire si les usines sont reclassées dans le secteur du commerce de gros. Dans le cas des bases de données de courte durée, les transferts sont considérés comme des entrées et des sorties et nous n'avons pas essayé d'en mesurer précisément l'importance.

Chevauchement des catégories d'entrée d'entreprises. La définition de l'entrée et de la sortie en termes de nombre d'usines pose peu de problèmes de chevauchement. Les usines sont classées dans une seule catégorie. Le problème de chevauchement est potentiellement plus sérieux lorsqu'on mesure l'entrée à l'aide du nombre d'entreprises. Une entreprise peut entrer dans l'industrie en construisant une usine, en faisant l'acquisition d'une à un état financier typique des revenus et dépenses d'une société. On l'envoie aux petits établissements manufacturiers dont la valeur des livraisons se situe en deçà d'un seul minimum" (Statistique Canada, 1979, page 10).

Certaines usines très petites ne reçoivent ni le questionnaire complet ni le questionnaire abrégé. Les données relatives à ces usines sont tirées des dossiers administratifs fiscaux plutôt que des questionnaires postaux. A la fin des années 1970 et au début des années 1980, les deux types de petits établissements[18] rendaient compte de 5 pour cent ou moins de toutes les livraisons du secteur manufacturier: 2.0 pour cent en 1970, 4.1 pour cent en 1975 et 3.4 pour cent en 1982.[19] Par contraste, ces établissements représentaient 40 pour cent de tous les établissements de ce secteur en 1970, 50 pour cent en 1975 et 53.9 pour cent en 1982.[20]

Il est important de comprendre la différence entre un grand et un petit établissement parce qu'il est parfois opportun, pour des raisons de coût, d'utiliser seulement un sous-ensemble de tous les établissements pour mesurer l'entrée et la sortie. De plus, la création et la disparition de petits établissements peuvent dépendre de l'empressement que l'on met à trouver ces petits établissements. Cet empressement peut aussi varier d'une année à l'autre en fonction des contraintes budgétaires auxquelles l'organisme statistique est soumis et des problèmes administratifs que pose le flot de paperasserie que l'on impose aux entreprises plus petites.

Dans le cadre de la présente étude, nous avons retenu essentiellement les établissements qui remplissaient un questionnaire complet pour les raisons invoquées plus haut et parce que l'échantillon visé par ce questionnaire nous permettait de mesurer d'une façon logique un plus grand nombre de caractéristiques des entrants. Cela est dû au fait que les questionnaires complets renferment des renseignements plus détaillés sur les activités des usines et que certains concepts, comme la valeur ajoutée, ne sont pas définis de la même manière dans le cas du questionnaire complet et du questionnaire abrégé.[21]

Nous avons examiné les répercussions de l'utilisation de cet échantillon en comparant les taux d'entrée et de sortie établis à l'aide de l'univers des établissements de recensement et ceux établis à l'aide de l'échantillon visé par le questionnaire complet seulement. Nous avons utilisé à cette fin la base de données de longue durée, 1970 représentant l'année initiale et 1979, l'année finale. L'échantillon visé par le questionnaire complet donne un taux d'entrée et de sortie beaucoup plus bas que l'échantillon intégral lorsque nous nous servons du nombre d'usines et d'entreprises; toutefois, son utilisation ne modifie pas beaucoup l'estimation de ces taux lorsqu'on les mesure en termes d'effectifs ou de livraisons.[22] Nous y reviendrons plus loin. Les petits établissements sont nombreux mais ne représentent qu'un faible pourcentage des effectifs totaux.

Les raisons qui nous ont amenés à ne retenir que le questionnaire complet pour les établissements nous ont aussi conduit à choisir les entreprises qui ne possédaient que les établissements visés par ce questionnaire pour les bases de données de courte durée. Aux fins du recensement des manufactures, une entreprise est constituée de tous les établissements qu'elle possède. On réserve spécifiquement les questionnaires complets aux établissements d'entreprises plus importantes et les questionnaires abrégés, aux petites entreprises.

L'utilisation du questionnaire complet crée un certain problème de dépendance. La démarcation entre un établissement visé par un questionnaire abrégé et celui visé par un questionnaire complet a changé radicalement en 1975. Cela ne pose pas de problèmes majeurs pour la base de données de longue durée. Ce ne serait pas le cas si on comparait les périodes 1970-75 et 1976-80, parce qu'il y aurait un peu moins d'entrants dans la dernière période. Le problème est plus aigu lorsqu'on mesure les taux annuels d'entrée et de sortie puisqu'une discontinuité prend de l'importance au milieu de la période. Nous traiterons de ce sujet dans une autre section.

* Unités de mesure

L'importance de l'entrée peut se mesurer en termes de nombre d'établissements et d'entreprises, ou d'extrants et d'intrants. Nous utilisons les deux ensembles de mesures. Les taux d'entrée et de sortie calculés à l'aide du nombre d'établissements et d'entreprises indiquent si le processus d'entrée et de sortie est facile; s'ils sont calculés à l'aide du volume d'extrants ou d'intrants, les taux déterminent l'importance de ce processus. Les deux variables des livraisons et des effectifs totaux (salariés) permettent de mesurer le volume. Les livraisons représentent la mesure la plus logique pour étudier le processus concurrentiel parce que celles-ci indiquent la part du marché que les entrants peuvent occuper. De son côté, la mesure des effectifs nous permet de connaître l'incidence de l'entrée et de la sortie sur la rotation du personnel.

La mesure des effectifs provient dans tous les cas des statistiques de l'activité totale obtenues à partir du recensement.[23] Dans le cadre de celui-ci, elle est présentée comme un équivalent annuel. Si, par exemple, une usine emploie 60 travailleurs par mois pour une période de six mois, on enregistrera 30 années- hommes. Dans certains cas, cette façon de procéder peut se traduire par une tendance à la baisse des estimations d'entrée et de sortie (puisque 60 personnes et non 30 sont touchées par la sortie de l'usine décrite ci-dessus). Elle modifierait également les taux d'entrée et de sortie calculés puisque ce facteur n'aura vraisemblablement pas la même incidence sur l'emploi dans les usines en exploitation, qui forment le dénominateur de ce calcul. On pourrait aussi supposer que les entrants et les sortants sont répartis uniformément sur toute l'année et que la durée de vie moyenne est de six mois. Tous les chiffres des effectifs bruts qui se rapportent à l'entrée et à la sortie seraient alors doublés.[24]

identificateurs d'entreprise qui ont été faites indiquent que les changements à cet égard tiennent compte de fluctuations économiques importantes. Il ne s'agit pas de simples changements de raison sociale, de réorganisations de société mineures ou d'erreurs de codage.

PROBLÈMES DE MISE EN APPLICATION: QUESTIONS GÉNÉRALES

Les grandes questions d'ordre conceptuel comme la période à étudier, le niveau de détail d'industrie et les catégories d'entrée et de sortie à choisir sont relativement simples à résoudre. Il en va tout autrement des problèmes relatifs aux particularités de chaque base de données qui rendent difficile la mesure précise des entrées et sorties. Nous donnons dans cette section une vue d'ensemble de certains de ces problèmes et de leur importance pour les données tirées du recensement des manufactures et utilisées dans ce document. Nous décrirons plus en détail dans les sections suivantes la solution propre à chaque problème et nous rendrons compte alors des difficultés posées par chaque base de données.

* Couverture

La valeur des statistiques d'entrée et de sortie dépendra de l'exhaustivité de la couverture offerte par la base de données. Des bases de données, comme celle de Dun and Bradstreet utilisée par Birch (1979, 1981), et la U.S. Small Business Administration (1984), sont incomplètes ayant été créées uniquement à partir des dossiers fournis par les sociétés qui voulaient être inscrites dans ces bases de données à des fins d'estimation de leur solvabilité. D'autres bases de données, comme celles qu'ont créées Storey[17] et ses collègues du R.-U., sont établies à partir de différentes sources, dont aucune ne peut se comparer à un recensement complet.

L'utilisation du recensement canadien des manufactures pour mesurer l'entrée et la sortie permet de surmonter en grande partie ces problèmes. Les données canadiennes embrassent toutes les entreprises du secteur manufacturier et sont recueillies par un organisme statistique officiel. Ces données sont le résultat combiné de l'expertise du personnel et de la vaste couverture que supposent la réalisation de recensements nationaux.

Les bases de données longitudinales peuvent aussi poser des problèmes, pas tellement en raison d'une couverture incomplète, mais parce qu'elles ne sont pas courantes ou qu'elles subissent des modifications. Cela est souvent dû au fait que l'on tarde à ajouter de nouvelles entreprises à une base de données ou à éliminer celles qui n'existent plus. De soudaines vagues d'activité pour saisir des entreprises existantes qu'on aurait peut-être oubliées ou pour éliminer des fichiers les entreprises défuntes, peuvent produire un faux niveau des entrées et des sorties mesurées pour une année donnée.

Ces problèmes ne se posent généralement pas pour le recensement des manufactures parce que celui-ci est annuel. Les dossiers administratifs fiscaux constituent un bon moyen de trouver de nouvelles usines et entreprises. De plus, un personnel qualifié est chargé de faire le suivi des entreprises qui existaient déjà et qui n'ont pas retourné leur questionnaire de recencement, pour vérifier si elles sont toujours en exploitation. En général, le recensement canadien renferme donc des données détaillées et courantes à partir desquelles on peut obtenir des taux annuels des entrées et des sorties significatifs. Certains retards et omissions peuvent se produire, mais ils seront négligeables comparativement à ceux d'autres sources.

Les données du recensement canadien ne sont pas complètement à l'abri d'un changement de couverture sur une période particulière. Un tel changement s'est produit au milieu des années 1970. Il existe toutefois des informations qui permettent d'estimer l'effet précis d'un changement de couverture du recensement.

Ce problème ne devrait pas avoir de répercussions sur la base de données de 1970- 79 qui a été créée pour mesurer l'entrée et la sortie sur une longue durée, puisqu'une grande partie des entrants oubliés au milieu des années 1970 auront été introduits dans la base avant 1979. Toutefois, il n'en a pas été ainsi pour les bases de données des établissements et des entreprises qui servent à mesurer les taux annuels d'entrée, et des modifications ont dû être apportées pour faire face à ce problème. Nous traiterons de ces bases dans une autre section.

* Choix de l'échantillon

L'avantage d'un recensement officiel repose sur la large couverture qu'il offre. Par contre, l'utilisation de tous les questionnaires pour faire l'analyse peut être très coûteuse. De plus, il faut se rappeler que tous les questionnaires ne sont pas d'égale qualité.

Les établissements interrogés directement par Statistique Canada dans le cadre du recensement annuel des manufactures peuvent recevoir un questionnaire complet ou abrégé. Voici quelle est la différence entre les deux:

"Le questionnaire complet est un questionnaire très détaillé envoyé aux établissements dont la valeur des livraisons dépasse des seuils minimums qui varient selon la province et l'industrie d'une année à l'autre; il est conçu pour prendre en compte à peu près toutes les livraisons de l'industrie. En 1975, seule une faible partie (4.1%) de la valeur des livraisons des produits de propre fabrication des industries manufacturières n'a pas figuré sur les questionnaires complets. La formule abrégée constitue un questionnaire simplifié qui ressemble beaucoup corresponde à de véritables créations et disparitions d'entreprises par suite de l'ouverture ou de la fermeture d'une usine, étaient plus grandes.

Afin d'être en mesure d'évaluer les types de changements qui se sont produits lors de l'apparition ou de la disparition d'un code ENT au moment de l'acquisition ou de l'aliénation d'une usine, tous les établissements visés ont été classés dans une catégorie ou plus. Cet exercice avait deux buts. Le premier consistait à évaluer l'importance des changements apportés à la base de données et attribuables à des acquisitions et à des aliénations, et le second, à isoler le nombre de cas où le changement de code ENT n'était dû qu'à des causes mineures, comme un changement de raison sociale qui n'était pas accompagné d'un événement important. Une cause mineure est plus difficile à définir qu'une cause majeure. Par conséquent, nous avons procédé par rétro-élimination, en faisant abstraction des cas de réorganisation importante et en examinant par la suite la catégorie restante.

Trois événements ont été jugés suffisamment importants pour écarter les changements organisationnels mineurs. Le premier événement correspondait au changement du pays contrôleur de l'entreprise à qui appartenait l'usine en 1970 comparativement à 1979.[13] Le deuxième événement consistait à vérifier si l'entreprise qui avait fait une acquisition ou aliéné une usine, avait poursuivi ses activités tout au long de la décennie.[14] Dans le premier cas, l'entreprise acquérante possédait en 1970 une usine qui faisait partie d'une industrie à quatre chiffres autre que celle dont faisait partie en 1979 l'usine acquise. Dans le second cas, l'entreprise qui avait quitté une industrie en aliénant une usine, pouvait se retrouver dans une autre industrie en 1979. Le troisième événement correspondait à une fusion horizontale.[15] Cela se produisait lorsqu'une entreprise qui entrait dans une industrie par suite de l'acquisition d'usines le faisait en acquérant des usines de plus d'une entreprise. Il est peu probable que l'un ou l'autre de ces événements importants ait pu avoir lieu sans un changement organisationnel majeur.

Chaque usine qui a fait l'objet d'une acquisition ou d'une aliénation a été classée en fonction des catégories d'événements importants. Comme les catégories ne s'excluent pas mutuellement, une usine pourrait être classée dans plus d'une catégorie. On mesure l'importance d'une catégorie pour les entrants selon le ratio de la somme des livraisons en 1979 de toutes les usines dans toutes les industries contenues dans cette catégorie, divisé par les livraisons en 1979 de toutes les usines acquises par les entreprises qui entrent dans une industrie. L'évaluation de l'importance des différentes catégories à partir des usines qui appartenaient à des entreprises qui ont quitté une industrie en aliénant celles- ci, se mesure de la même façon, mais en utilisant les livraisons de 1970.

Près de la moitié des livraisons de 1979 des usines acquises lors de l'entrée des entreprises était attribuable à des usines acquises par des entreprises qui possédaient des usines dans une autre industrie canadienne en 1970. Il s'agissait alors de fusions aux fins de diversification et non d'une forme mineure de réorganisation de société. Le changement du pays contrôleur a eu une influence sur environ 43 pour cent des livraisons des usines acquises par de nouvelles entreprises. Les fusions horizontales au sein de la catégorie des acquisitions et des aliénations ont été moins importantes. Environ 13 pour cent des livraisons en 1979 provenaient d'usines qui avaient été fusionnées avec d'autres usines de la même industrie à un moment donné au cours de la décennie, et qui s'inscrivaient dans le processus d'acquisition qui avait intégré de nouvelles entreprises dans une industrie. Après avoir pris en considération tous les événements principaux, nous avons constaté que seulement 25.3 pour cent des livraisons n'étaient pas rattachés à une acquisition par une entreprise canadienne existante, à un changement de pays contrôleur ou à une fusion horizontale.

Nous avons pu observer, en faisant une analyse semblable des sorties, que le pourcentage des cas d'aliénation qui n'étaient pas accompagnés d'une fusion horizontale, d'un changement de pays contrôleur ou de la continuation dans une autre industrie des activités de l'entreprise aliénante, était quelque peu supérieur à celui des entrées (environ 35.1 pour cent).

Les deux ensembles d'usines dans chacune des catégories restantes ne se chevauchent pas entièrement. L'application simultanée des événements principaux aux usines acquises et aux usines qui ont fait l'objet d'une aliénation, révèle que seulement 8.6 pour cent de toutes les usines qui ont enregistré 9.5 pour cent des livraisons restantes n'ont peut-être pas subi une réorganisation majeure.

Nous avons fait une vérification manuelle des usines de la dernière catégorie. Il en est ressorti qu'un changement mineur dans l'état de l'entreprise, tel un changement de raison sociale, était intervenu pour 3.4 pour cent des établissements originaux enregistrant 1.6 pour cent des effectifs. Comme on peut le constater dans les Tableaux 3 et 4, la reclassification dans la catégorie des établissements en exploitation (15) du groupe qui faisait partie des catégories de l'acquisition (22) et de l'aliénation (31), n'a eu aucun effet sur l'importance de ces catégories.

Nous pouvons conclure qu'un examen des différents types d'entrée du fait d'une acquisition et de sortie du fait d'une aliénation, confirme la désignation de cette catégorie. Compte tenu du mode d'attribution des identificateurs de nouvelles entreprises utilisé dans beaucoup de bases de données, il peut toujours arriver que le phénomène que l'on mesure ne soit pas associé à un changement important dans la structure de contrôle ou d'exploitation du groupe.[16] Les réorganisations de société qui donnent lieu à une nouvelle structure juridique sans changer la structure ou les principes directeurs en matière de propriété ou d'exploitation peuvent se produire pour plusieurs raisons (des réductions d'impôts, par exemple). Les vérifications de validation des comparé les variables de la raison sociale, de la propriété et de l'emplacement de toutes ces usines avec celles de nouvelles usines pour l'année 1985 pour voir s'il y avait des correspondances avec une des usines qui avait disparu. Parmi les usines en exploitation en 1981, 12,076 avaient disparu en 1985. Ces usines employaient 206,668 travailleurs en 1981. De ce nombre, seulement 10 usines employant 209 employés pouvaient avoir été l'objet d'une erreur de codage. Nous n'avons pu trouver que 10 nouvelles usines en 1985 qui semblaient correspondre à une usine de 1981 qui avait perdu son numéro d'identification. Cela nous amène à supposer que la surestimation du taux de fermeture, établi en fonction de la disparition de l'identificateur de l'usine, est négligeable.

Des problèmes administratifs peuvent également causer le type d'erreur opposé. Si l'on attribue de nouveaux identificateurs d'usine lorsqu'on ne le devrait pas, il en résulte une sous-estimation du taux de sortie en 1981 et du taux d'entrée en 1985. Nous avons aussi examiné la fréquence du type d'erreur opposé. En effet, nous avons vérifié si la base de données renfermait deux identificateurs d'usine identiques pour les années 1981 et 1985, lorsque les trois variables (raison sociale, propriété, emplacement) avaient changé. Les codes d'identification de 1981 de ces usines auraient dû être changés et remplacés par de nouveaux codes avant 1985. Nous avons trouvé environ 18 usines employant au total 1,298 travailleurs. Encore une fois, le nombre d'usines de cette catégorie représentait un pourcentage insignifiant du total des sorties.

Le premier type d'erreur et le second type d'erreur occasionneraient respectivement une hausse et une baisse constantes des taux de création et de disparition des usines. Dans chaque cas, les erreurs étaient minimes et essentiellement neutralisantes. De plus, le nombre d'erreurs était probablement gonflé puisque l'identification des problèmes potentiels de codage reposait sur les programmes informatiques et que l'on ne s'est pas attardé aux erreurs réelles de codage étant donné que le taux d'erreur potentiel maximal était déjà très bas.

Compte tenu de la nature des critères utilisés pour réattribuer l'identificateur d'usine NSD et du soin apporté par Statistique Canada au respect de ces critères, nous pouvons conclure que, dans le cadre du recensement annuel des manufactures, l'apparition des nouveaux codes d'établissement et la disparition des anciens peuvent généralement être attribuées à de "véritables" ouvertures et fermetures. Cela n'est pas le cas de certaines autres bases de données où un changement de personne juridique est une justification suffisante pour supprimer un code et en attribuer un autre.[12] Dans cette étude, la propriété et la raison sociale de l'usine peuvent changer, mais dans la mesure où l'emplacement reste le même, il n'y aura ni changement d'identificateur ni fausse indication quant à l'ouverture et à la fermeture d'une usine.

* Le code de l'entreprise

Les identificateurs d'entreprise (codes ENT) ont permis de suivre l'évolution des groupes d'établissements sous contrôle commun. Le même identificateur ENT a été attribué à toutes les usines de l'industrie manufacturière, forestière et minière appartenant à la même entreprise. Ce code ne représente pas la personne juridique, mais s'apparente intentionnellement au concept d'une entreprise analysé plus haut. Les personnes juridiques sont déjà identifiées par des codes (S.I.R.E.); en cas de changement de personne juridique (à la suite d'une incorporation, d'une fusion ou d'une réorganisation des établissements) ou de propriétaire, de nouvelles valeurs sont attribuées aux codes S.I.R.E. et les anciennes sont supprimées. Etant donné que l'identité de la personne juridique change beaucoup plus souvent que l'entreprise qui contrôle la personne juridique, l'utilisation d'un code (S.I.R.E.) pour identifier celle-ci peut produire de "fausses" créations et disparitions. Dans nos diverses études, les créations sont considérées comme fausses si elles ne supposent que des changements mineurs qui n'entrent ni dans la catégorie des entrées par suite de la construction d'une usine ni dans celle des entrées relatives à l'acquisition que nous avons défini précédemment.

Par contre, des changements de code ENT dans la base de données ne rendent fondamentalement compte que des transformations importantes dans l'organisation d'une entreprise. L'apparition d'un code ENT dans une industrie devrait correspondre à une entrée à la suite de la construction ou de l'acquisition d'une usine (où l'acquisition englobe des changements de contrôle qui n'aboutissent pas nécessairement à la fusion des installations de l'entreprise acquise avec celles de l'acquéreur). De la même façon, la suppression d'un code ENT devrait correspondre à une fermeture. Comme dans le cas des établissements, on ne devrait pas supprimer les codes des entreprises qui sont toujours en exploitation et en attribuer de nouveaux sans avoir une bonne raison de le faire. Toutefois, contrairement aux établissements, les règles qui régissent la réattribution des codes des entreprises en exploitation ne sont pas aussi clairement définies. Cela s'explique en partie par la plus grande complexité des événements dont devrait tenir compte toute définition. La règle relative à la raison sociale, à l'emplacement et à la propriété, qui s'applique au changement d'identificateur d'une usine ne serait pas suffisante.

Les codes ENT ne devraient être changés que lorsqu'un événement important se produit dans la vie de l'entreprise. Nous avons examiné attentivement dans quelle mesure cela se transpose dans la base de données. Pour ce faire, nous n'avions pas à vérifier toutes les catégories qui comportaient une attribution ou une suppression de code ENT. Puisqu'un code NSD et un code ENT avaient été attribués à chaque établissement, nous avons porté notre attention exclusivement sur les entreprises qui ont fait leur entrée dans une industrie en acquérant une usine ou qui ont quitté l'industrie en aliénant une usine (catégories 22 et 31 du Tableau 1). De cette façon, nous éliminions les cas de création ou de disparition d'entreprises par suite de l'ouverture ou de la fermeture d'un établissement (catégories 23 et 34). Compte tenu du soin apporté par Statistique Canada à l'attribution et à la suppression des codes d'établissement, les chances que le dernier ensemble d'événements

"Il existe en fait un niveau intermédiaire entre l'établissement et l'entreprise: la personne juridique. Il s'agit de l'unité de propriété. Les personnes juridiques peuvent être constituées ou non en sociétés par actions, ou encore être formées d'individus. Une personne juridique peut en posséder une autre, de sorte qu'il est possible pour une entreprise de contrôler plusieurs personnes juridiques, tout comme une personne juridique peut posséder plusieurs unités d'exploitation (établissements)" (Statistique Canada, 1983, page 24).[6]

Puisqu'une entreprise correspond à l'unité qui regroupe tous les établissements en propriété commune, on peut créer des sous-unités qui combinent tous ces établissements dans un regroupement d'industries particulier (l'industrie à deux, à trois ou à quatre chiffres). Ainsi, on peut mesurer l'entrée des entreprises au niveau d'une industrie particulière ou pour l'ensemble du secteur manufacturier. La base de données de longue durée (1970-79) contient des informations au niveau de détail le plus fin (quatre chiffres). Pour mesurer les taux annuels d'entrée et de sortie, on a recours au secteur manufacturier dans son ensemble. Lorsqu'on compare les résultats d'un ensemble de données à ceux d'un autre, il faut se rappeler qu'il n'est pas nécessaire d'obtenir les mêmes estimations pour les raisons exposées plus haut.

DÉTERMINATION DE L'ANNÉE D'ENTRÉE ET DE SORTIE

On attribue à chaque établissement un numéro d'identification unique, le numéro de série de dossier (NSD).[7] L'établissement conserve ce numéro aussi longtemps qu'il est inclus dans le recensement des manufactures.[8] On attribue également un identificateur unique à chaque entreprise (que l'on appellera ici le code ENT).[9] Contrairement au NSD d'un établissement, le code d'entreprise peut changer lorsqu'une entreprise en achète une autre.

La création d'une usine ou d'une entreprise correspond à l'apparition d'un nouveau code d'identification, et une sortie se produit lorsque le code disparait. Si le code demeure en vigueur tout au long de la période qui fait l'objet de l'étude, on dit que l'usine ou l'entreprise est en exploitation. La base de données de courte durée se sert d'années contigues entre 1970 et 1982 pour comparer l'état des établissements. La base de données de longue durée compare l'état de ceux-ci la première et la dernière année de la décennie 1970.

Ce sont, premièrement, l'état (la continuation des opérations, l'interruption des opérations, ou la création) d'une usine et, deuxièmement, le niveau d'activité qui permettent de définir l'entrée et la sortie. Le nombre d'entrants est calculé la première année où l'identificateur fait son apparition et que le chiffre des effectifs ou la valeur des livraisons du fabricant est positif; on considère que les sorties se produisent l'année qui précède la disparition de l'identificateur, ou l'année même de cette disparition lorsque le chiffre des effectifs ou la valeur des livraisons tombe à zéro.[10] Le dernier critère permet d'exclure du compte des sorties les unités de production qui, pour une raison quelconque, avaient déjà cessé leurs activités.

VALIDATION DES IDENTIFICATEURS

L'examen des changements apportés aux identificateurs d'entreprise et d'établissement permet de mesurer l'entrée et la sortie. Nous verrons dans cette section pourquoi ces identificateurs apparaissent et disparaissent.

* Le code de l'établissement

On considère que l'entrée et la sortie d'usines coincident avec l'apparition et la disparition d'un code d'établissement (le NSD). Que cette définition produise des estimations significatives des ouvertures et des fermetures d'établissement dépend de la méthode utilisée par l'organisme statistique pour attribuer les codes d'établissement. La fermeture d'un établissement justifie habituellement le retrait d'un code; toutefois, il peut arriver que l'on réattribue à des usines en exploitation des codes NSD: on remplace l'ancien code par un nouveau. Si cela se produit, il y aura surestimation des ouvertures et des fermetures.

Des problèmes se posent à ce chapitre parce que les établissements, comme les entreprises, comportent plusieurs caractéristiques. Certaines de ces caractéristiques peuvent changer pendant la durée de vie d'un établissement et amener le système de codage administratif à attribuer un nouveau numéro même si l'établissement n'a pas cessé ses activités. Si, par exemple, l'une d'elles change (tel la propriété) entraînant une réattribution de code, la création et la disparition ne correspondra pas à une ouverture et à une fermeture d'établissement.

La signification de la création et de la disparition d'un établissement dépend donc du type d'événement qui a amené l'organisme statistique à attribuer un nouveau code NSD aux usines qui n'ont pas cessé leurs activités. La ligne de conduite de Statistique Canada consiste, dans le cas d'un établissement en exploitation, à rejeter le code et à en attribuer un autre seulement si les trois variables de l'emplacement, de la propriété et de la raison sociale changent simultanément.[11] Cette ligne de conduite écarte du compte des disparitions les cas où seulement la propriété ou la raison sociale change.

La validité des mesures d'entrée qui sont établies tient au degré de diligence avec lequel Statistique Canada applique cette ligne de conduite. Deux tests ont permis d'en faire la vérification. Premièrement, toutes les usines qui étaient en exploitation en 1981 et avaient disparu en 1985 ont fait l'objet d'un examen. Nous avons industrie à quatre chiffres, d'une entreprise qui n'était rattachée à aucune autre usine de cette industrie en 1970. Il y a disparition d'une entreprise lorsque l'entreprise qui était rattachée à une usine en 1970 n'était plus rattachée à aucune usine d'une industrie de quatre chiffres particulière à compter de 1979. L'entreprise a été définie comme tous les établissements sous contrôle commun au sein d'une industrie à quatre chiffres de la CAE. On prendra note que les usines ou les entreprises qui ont fait leur entrée dans une industrie après 1970 et qui ont disparu avant 1970 ne sont pas prises en compte dans les mesures d'entrée ou de sortie tirées de la base de données de longue durée.

En raison du lien qui unit l'usine et l'entreprise dans la base de données de longue durée, on peut mesurer plusieurs catégories d'entrée différentes. La matrice de l'état des usines et des entreprises présentée dans le Tableau 1 offre un résumé de ces catégories. Les codes d'identification de cellule, qui serviront par la suite à indexer les variables, sont également indiqués dans le tableau. On peut mesurer l'importance des différentes catégories en utilisant le nombre d'établissements, d'entreprises, de livraisons, d'emplois ou toutes autres variables contenues dans le recensement des manufactures.

Bases de données manufacturières annuelles des établissements et des entreprises

La deuxième et la troisième base de données permettent de suivre séparément et de façon longitudinale l'historique des entreprises et des établissements pour chaque année entre 1970 et 1982. Elles servent surtout à faire des comparaisons à court terme. Ces deux bases de données définissent l'entrée et la sortie en ce qui a trait au secteur manufacturier dans son ensemble. Les créations d'établissement y sont définies comme l'apparition de nouvelles usines dans le secteur manufacturier. Il en va de même pour la définition des nouvelles entreprises. Les usines ou les entreprises qui passent d'une industrie manufacturière de quatre chiffres à une autre ne sont pas considérées comme des entrants aux fins de ces bases de données. Par contre, celles qui faisaient partie d'un autre secteur (commerce de gros, par exemple) et qui entrent dans le secteur manufacturier sont considérées comme des entrants.

Le modèle de classification des entrées et des sorties utilisé pour ces bases de données est un peu moins détaillé que celui qui sert à l'analyse des données de longue durée. La classification de la base de données des établissements était triple: on y retrouvait les créations, les fermetures et les établissements en exploitation d'année en année. On ne tenait pas compte dans cette base de données de l'entreprise propriétaire: que l'usine ait été acquise ou aliénée n'était pas pris en considération. Le modèle de classification de la base de données des entreprises était plus détaillé. Au chapitre de l'entrée, une distinction était faite entre les nouvelles entreprises du fait de la construction d'une usine et les nouvelles entreprises du fait de l'acquisition d'une usine; on a fait une distinction semblable pour la sortie.

Les bases de données annuelles des établissements et des entreprises nous permettent de répondre à différentes questions. Dans le cas de la base de données des établissements, l'entrée se définit comme la création d'un établissement, et la disparition, comme la fermeture de celui-ci. Etant donné que cette base de données considère toutes les ouvertures et les fermetures d'usine pour les entreprises en exploitation, les entreprises nouvelles et celles qui disparaissent, elle permet d'évaluer dans quelle mesure la rotation des usines entraîne des changements dans l'emploi. La base de données des entreprises offre la possibilité de faire la distinction entre une entrée attribuable à des ouvertures et une entrée attribuable à des acquisitions. Elle permet également de comparer l'activité des nouvelles entreprises avec celles des entreprises déjà en exploitation. Cette base de données peut aussi servir à faire des distinctions semblables pour la sortie. On peut y avoir recours pour répondre à des questions concernant la dynamique du processus concurrentiel au niveau de l'entreprise.

Afin de pouvoir mieux comprendre la signification des mesures d'entrée et de sortie fournies par les trois bases de données, il faut examiner la définition des termes "établissements" et "entreprises" qui ont été utilisés et décrire plus en détail les différentes catégories choisies. C'est ce que nous ferons dans les deux sections suivantes.

DÉFINITION DES ÉTABLISSEMENTS ET DES ENTREPRISES

Pour mesurer l'entrée et la sortie, on se sert de deux unités de production de base: d'une part, l'établissement ou l'usine, qui signifient la même chose aux fins de ce document, et d'autre part, l'entreprise. Chaque terme doit être défini avec soin si l'on doit comparer les données canadiennes à celles d'autres pays autant qu'à celles d'autres ensembles de données du Canada.

Un établissement, selon Statistique Canada, correspond habituellement à une usine ou à un atelier.[2] Seuls les établissements qui font partie du secteur manufacturier[3] ont été retenus dans le cadre de ce document. L'établissement est l'unité statistique de base à partir de laquelle des données sont recueillies pour le recensement annuel des manufactures.[4]

Une entreprise est définie comme l'ensemble des établissements du secteur manufacturier sous contrôle commun.[5] Une entreprise est donc un concept qui ne correspond pas nécessairement à la personne juridique ou à ce que l'on désigne parfois par le terme entité commerciale ou morale. Statistique Canada résume ci- après la corrélation entre la personne juridique, l'établissement et l'entreprise:

entreprises en exploitation de celles qui sont nouvelles et qui disparaissent, afin de mesurer la contribution relative de chaque groupe à la hausse et à la baisse de l'emploi. A cette fin, il faut distinguer les entrées, les sorties et les entreprises en exploitation, de la création et de la disparition des usines. Cette ventilation est également requise pour étudier le processus concurrentiel. De plus, on aura avantage à ajouter une autre catégorie à ce type d'étude, soit l'entrée et la sortie d'entreprises par l'acquisition et l'aliénation d'une usine.

* Solution

Les différentes questions qui ont été soulevées dans cette section ont été résolues en adoptant un ensemble de mesures qui tiennent compte d'une longue et d'une courte durée. Les données agrégées comme les données désagrégées de l'entrée et de la sortie sont utilisées. Les données agrégées offrent un aperçu des taux annuels ou de courte durée des entrées et sorties d'établissement; elles permettent aussi de mesurer l'entrée de nouvelles entreprises par suite de la construction d'une usine. Au niveau de l'agrégation, ces données fournissent une indication raisonnable de l'activité totale au niveau d'industrie sous-jacent. Les données désagrégées servent aux estimations de longue durée. Ce sont celles qui offrent le plus de détails pour les estimations de longue durée et qui prennent en compte la valeur cumulative du changement. Les établissements et les entreprises sont liés ensemble pour que l'on puisse mesurer l'activité de création et de fusion des usines par les entreprises nouvelles et en exploitation.

LES BASES DE DONNÉES

Puisqu'il fallait mesurer l'entrée sur différentes périodes (longue v. courte durée), pour différents niveaux d'agrégation des industries (secteurs particuliers par opposition au secteur manufacturier) et pour différentes unités de production (entreprises v. usines), trois différentes bases de données longitudinales ont dû être créées. Ensemble, ces trois bases de données permettent de mesurer l'entrée et la sortie pour une industrie en particulier et le secteur manufacturier en général, entre deux années contigues ou sur une plus longue période, à l'aide d'entreprises et d'établissements -- considérés isolément et ensemble.

Tableau I

Matrice de classification des usines et des entreprises utilisée pour étudier l'entrée et la sortie du secteur manufacturier canadien

	Etat de l'entreprise		
Etat de l'usine	En exploitation	Nouvelle	Fermée
Aliénations	11	s.o.	31
Acquisitions	12	22	s.o.
Créations	13	23	s.o.
Disparitions	14	s.o.	34
En exploitation	15	s.o.	s.o.
Transfert interne	16	26	s.o.
Transfert externe	17	s.o.	37

Définitions	Cellule	
Entrants	22	Entreprises qui ont fait leur entrée dans l'industrie en acquérant une usine ou plus entre t et t + n
	23	Entreprises qui ont fait leur entrée dans l'industrie en ouvrant une usine ou plus entre t et t + n
	26	Entreprises qui ont fait leur entrée dans l'industrie en transférant une usine ou plus d'une
Sortants	31	Entreprises qui ont quitté l'industrie en aliénant une usine ou plus entre t et t + n
	34	Entreprises qui ont quitté l'industrie en fermant une usine ou plus entre t et t + n
	37	Entreprises qui ont quitté l'industrie en transférant une usine ou plus de l'industrie donnée à une autre industrie entre t et t + n
En exploitation	11	Entreprises en exploitation qui ont aliéné une usine ou plus entre t et t + n
	12	Entreprises en exploitation qui ont acquis une usine ou plus entre t et t + n
	13	Entreprises en exploitation qui ont construit une usine ou plus entre t et t + n
	14	Entreprises en exploitation qui ont fermé une usine ou plus entre t et t + n
	15	en exploitation qui possédaient au moins une usine pendant la période t et la période t + n
	16	Entreprises en exploitation qui ont transféré des usines dans l'industrie donnée
	17	Entreprises en exploitation qui ont transféré des usines à l'extérieur de l'industrie donnée

*s.o. = sans objet

La base de données de longue durée de l'industrie à quatre chiffres

La première base de données mesure l'entrée et la sortie sur une longue durée en comparant l'état des unités de production en 1970 et en 1979. Elle fournit des détails sur l'état de l'établissement et de l'entreprise et lie les deux. Elle peut donc servir à mesurer l'entrée et la sortie d'usines et d'entreprises. Cette base de données permet également de mesurer la rotation des usines appartenant à des entreprises en exploitation de façon à offrir une norme de comparaison pour le secteur des entreprises qui entrent dans une industrie ou la quittent. Finalement, elle mesure l'activité au niveau plus détaillé de l'industrie à quatre chiffres de la Classification des Activités Économiques (CAE). Une création d'usine a été définie comme l'apparition en 1979, dans une industrie à quatre chiffres, d'une usine qui ne faisait pas partie de cette industrie en 1970. Une fermeture d'usine correspondait à la disparition en 1979 d'une usine qui faisait partie d'une industrie à quatre chiffres et qui existait en 1970. On a défini une entrée comme l'apparition en 1979, dans une

1990). Nous décrirons dans les sections suivantes la nature des problèmes d'ordre conceptuel qu'il a fallu résoudre, les bases de données utilisées et les catégories choisies.

QUESTIONS D'ORDRE CONCEPTUEL

Des décisions doivent être prises en ce qui concerne le bon niveau d'agrégation des industries, l'unité de production et les catégories d'entrée et de sortie à utiliser, ainsi que la période choisie pour mesurer l'entrée et la sortie. Les choix appropriés pour chacune de ces questions sont intimement liés.

* Choix du niveau d'industrie

On peut mesurer l'entrée et la sortie au niveau de l'agrégation des industries du secteur manufacturier dans son ensemble ou à un niveau plus fin comme celui d'une industrie de la CAE à quatre chiffres. Puisque les économistes industriels cherchent habituellement à savoir dans quelle mesure l'entrée et la sortie facilitent le processus équilibrant, il faut obtenir des statistiques au niveau de chaque industrie.

Néanmoins, des statistiques d'entrée et de sortie agrégées peuvent être utiles. Premièrement, il peut être intéressant de savoir combien d'étrangers au secteur manufacturier ont fait leur entrée dans ce secteur. Deuxièmement, lorsque les données agrégées sont représentatives des données d'un niveau d'industrie particulier, elles résument bien les tendances sous-jacentes et ce, à un coût beaucoup plus avantageux que celui occasionné par la création d'une série de données pour chaque industrie. Lorsqu'on mesure l'entrée des usines, les données agrégées représentent d'une façon satisfaisante le nombre d'entrées totales dans chaque industrie. Cela vient du fait qu'une nouvelle usine qui entre dans une industrie manufacturière donnée à quatre chiffres est également un entrant pour le secteur manufacturier dans son ensemble. Les taux agrégés de création et de disparition d'usines représentent donc un moyen potentiellement bon de résumer l'activité sous-jacente au sein de chaque industrie.

L'utilité des taux agrégés d'entrée et de sortie des entreprises est plus problématique. Il n'y a pas nécessairement autant d'entreprises qui entrent dans le secteur manufacturier que d'entreprises qui entrent dans chaque industrie à quatre chiffres. Une entreprise peut entrer dans une industrie particulière à quatre chiffres sans être considérée comme un entrant dans le secteur manufacturier en général -- si elle existait déjà dans une autre industrie à quatre chiffres. Les taux agrégés d'entrée des entreprises permettront de mesurer la somme d'activité au niveau d'industrie sous-jacent, si la plupart des entrées au niveau de chaque industrie à quatre chiffres sont le fait d'entreprises qui étaient étrangères à cette industrie ainsi qu'au secteur manufacturier en général. Que ce soit ou non le cas relève du domaine empirique.

* Données de l'entreprise v. données de l'établissement

Les économistes industriels s'intéressent au processus de rotation des entreprises et des usines pour les répercussions que celui-ci peut avoir pour une période donnée sur l'évolution des profits de l'industrie, l'innovation et la productivité. De telles considérations laissent supposer que l'unité d'analyse appropriée est l'entreprise plutôt que l'unité de production particulière (l'usine ou l'établissement). C'est l'entreprise, et non l'usine, qui décide d'entrer dans une industrie ou de la quitter.

D'autre part, les taux d'entrée des usines sont utiles à connaître puisqu'ils donnent une vue d'ensemble de l'importance de <u>toutes</u> les nouvelles usines qui sont créées par les entreprises qui entrent dans une industrie et par celles qui sont déjà en exploitation. Cette variable, dont l'influence sur le processus équilibrant entraîne vers le bas les profits supra-normaux, peut être des plus déterminante. Dans le cas des études de création d'emplois, c'est le processus d'ouverture et de fermeture des usines, plutôt que le processus de création d'entreprises (qui comporte également une composante de fusion) qui est pertinent. Il ressort de tout cela que les mesures d'entrée et de sortie qui prennent en compte l'activité des entreprises et des établissements sont utiles dans différents contextes.

* Période

Il faut choisir la durée sur laquelle portera la mesure de l'entrée et de la sortie. On peut l'estimer en comparant, à l'aide de données annuelles, l'état des entreprises et des usines à deux points contigus dans le temps, ou en utilisant des points limites qui sont les plus éloignés l'un de l'autre. Dans le dernier cas, on ne tient pas compte de l'état des entreprises dans l'intérim. Le premier moyen permet de mesurer les taux de changement annuels; l'autre examine l'effet cumulatif de l'entrée et de la sortie sur la période donnée. Ces deux mesures, considérées ensemble, peuvent servir à évaluer l'importance de l'entrée dans le processus concurrentiel et à déterminer jusqu'à quel point les statistiques de rotation du personnel de courte durée rendent surtout compte d'in phénomème transitoire ou de plus longue durée.

* Catégories d'entrée et de sortie

Le type de recherche détermine également la nature des catégories d'entrée et de sortie à utiliser. Pour faire des études de la rotation du personnel, il faut pouvoir compter sur des systèmes de classification qui mettent l'accent sur l'entrée et la sortie des unités de production réelles, soit les usines ou établissements. Dans ces études, la distinction entre entreprises et usines doit être faite. Il est également important de séparer les

MÉTHODE UTILISÉE POUR MESURER L'ENTRÉE ET LA SORTIE DU SECTEUR MANUFACTURIER CANADIEN

Nous pouvons définir l'entrée et la sortie comme l'émergence de nouvelles unités de production et la disparition de vieilles unités. Malheureusement, s'il est relativement facile d'en donner une définition, il demeure difficile de mesurer précisément l'entrée et la sortie. On peut interpréter de plusieurs façons les termes "nouvelle" et "vieille". Il se peut que les données ne puissent pas se prêter à l'évaluation des concepts désirés. Il s'ensuit que pour faire un travail empirique dans ce domaine, il faut préciser avec soin les concepts à mesurer et les méthodes à utiliser.

Nous n'avons absolument pas tenu compte de tous ces points dans la section précédente où nous présentions un ensemble de résultats sur l'importance de l'entrée et de la sortie. Nous décrirons dans le reste de ce document la méthode utilisée pour générer des statistiques d'entrée et de sortie relatives au secteur manufacturier pour les années 1970 et le début des années 1980. Ce document s'inscrit dans une série de documents qui portent sur la dynamique du changement dans l'industrie canadienne. Il ne renferme pas de résultats détaillés, mais nous avons inclus certains extraits à des fins explicatives.

Pour pouvoir mesurer l'entrée et la sortie, il faut répondre à plusieurs questions. Premièrement, à quel type d'étude serviront les mesures d'entrée et de sortie? Bien que les bases de données dont il est question dans ce document aient été créées essentiellement pour les études du processus concurrentiel, elles ont aussi été utilisées pour étudier les changements au chapitre des emplois (Baldwin et Gorecki, 1989f, 1989g, 1990). Les mesures qui sont utiles à un type d'étude ne le sont pas nécessairement à un autre. Deuxièmement, il faut répondre à des questions plus générales: quelle période et quels types d'entrée et de sortie choisir, et quel devrait être le niveau de détail d'industrie? Troisièmement, quels problèmes se posent pendant le processus de mesure même?

Nous expliquerons d'abord comment la recherche qui comporte différents objectifs peut nécessiter différentes mesures. Nous examinerons ensuite le choix de la période, la définition de l'industrie et les catégories d'entrée. Suivront un aperçu général des bases de données ainsi qu'une définition des catégories d'entrée et de sortie réellement utilisées. Finalement, nous étudierons en détail les problèmes posés par la mise en application.

RAPPORT ENTRE LES DEFINITIONS ET LES OBJECTIFS

Les utilisateurs de données administratives et d'enquêtes doivent procéder avec prudence lorsqu'ils ont recours à ces sources pour des fins auxquelles elles n'étaient pas destinées au départ. Cela est particulièrement vrai lorsque l'attribution et la suppression des codes d'identification de ces bases de données servent à définir les créations et les disparitions. Les codes d'identification peuvent être attribués ou supprimés pour toutes sortes de raisons, et il se peut qu'aucune de ces raisons ne réponde à la définition particulière de l'entrée et de la sortie du chercheur.

Il y a beaucoup de façons de définir une création et une disparition puisqu'on se sert d'un vecteur de caractéristiques et non d'une seule dimension pour définir une entreprise. Au nombre de ces caractéristiques, mentionnons des variables comme l'industrie, la propriété, le pays contrôleur, la taille ainsi que l'emplacement et le nombre d'usines. La nature multidimentionnelle des caractéristiques de l'entreprise serait sans importance si une seule de ces caractéristiques était requise pour définir les créations et les disparitions, ou si elles changeaient toutes simultanément. Ce n'est pas le cas.

La nature de la recherche détermine la définition requise d'une nouvelle entreprise ou d'une entreprise qui disparait. Si le but de la recherche est de saisir comment la création d'entreprises influe en premier lieu sur le nombre d'emplois et la rotation du personnel, la définition d'une entreprise nouvelle est donc la plus appropriée, soit celle qui fait état de l'apparition d'une nouvelle entreprise suite à la construction d'une usine. Cette définition se fonde essentiellement sur les variables de l'état de l'usine et de l'emploi comprises dans le vecteur de caractéristiques qui définissent une entreprise. Une nouvelle entreprise est celle qui construit une usine, créant par le fait même de l'emploi. Une nouvelle entreprise qui n'est qu'une métamorphose d'une ancienne entreprise sous une nouvelle raison sociale ne devrait pas être définie comme une création aux fins des études qui portent sur la rotation du personnel. Par conséquent, il faut exclure les fusions de la définition d'une nouvelle entreprise qui sert à mesurer la rotation du personnel ou faire en sorte de les étudier séparément.[1]

Les études du processus concurrentiel requièrent une définition différente de l'entrée et de la sortie. Si la recherche vise à évaluer l'effet de la création d'une entreprise sur la concurrence, il convient alors de définir les entrées comme la création d'entités. Dans ce cas, on considérera les entrées issues de la construction d'une toute nouvelle usine (une catégorie qui dépend de la variable de l'état de l'usine comprise dans le vecteur de caractéristiques de l'entreprise) ainsi que les entrées présentes à la suite de l'acquisition d'usines existantes (une catégorie qui dépend de la variable de l'état de la propriété comprise dans le vecteur des caractéristiques de l'entreprise). Il y a lieu de faire la distinction entre les deux formes d'entrée, parce que leurs effets sur la performance ne seront pas nécessairement les mêmes.

La grande diversité d'interprétations que l'on peut donner à la notion d'entrée et de sortie indique qu'il est difficile de produire une seule estimation qui répond à plus d'une fin. Plusieurs bases de données ont donc été créées pour le travail qui fait l'objet des documents de recherche de Baldwin et de Gorecki (1989a à 1989g, et la cohorte. Inversement, la part moyenne de la valeur ajoutée augmente tout au long de la période (quelque dix années) étudiée ici. Le taux de croissance des entrants qui restent en exploitation représente plus qu'un effet de contrepoids au taux élevé de fermetures de chaque cohorte au cours des premières années de son existence.

La part moyenne de la valeur ajoutée d'une cohorte ainsi que les effets cumulatifs des cohortes successives sont représentés dans la Figure 2. La part moyenne du marché, établie à l'aide de la valeur ajoutée, de chaque cohorte d'entrants pour la période 1970-71 à 1980-81, sert de point de départ. La trajectoire de la part moyenne est ensuite appliquée à chaque cohorte. La part du marché total occupée par les entrants que l'on obtient est une représentation de l'effet cumulatif de l'entrée en moyenne. Au cours de la décennie étudiée, on ne constate pas de tendance à la baisse dans une part de cohorte moyenne et, par conséquent, l'effet cumulatif de l'entrée augmente continuellement. En dépit du taux élevé de fermetures, le nombre d'entrants qui restent en exploitation est suffisant pour qu'ils soient considérés comme un groupe.

La part de marché cumulative des entrants selon l'ouverture d'usines: une représentation

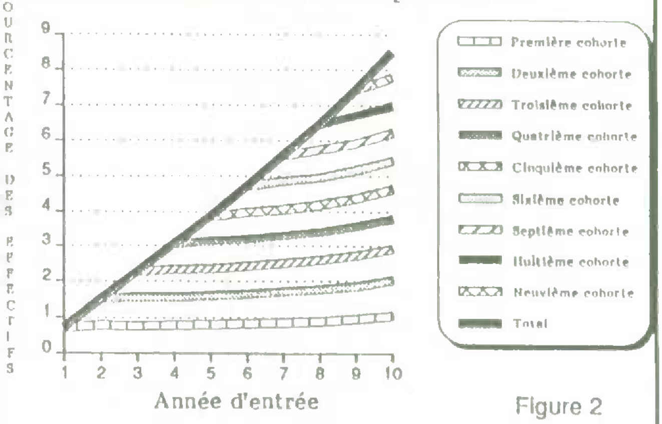


Figure 2

Effets cumulatifs de l'entrée et de la sortie

L'examen des taux d'entrée et de sortie de plus longue durée du secteur manufacturier canadien se fonde sur deux périodes de six années (1970-71 à 1975-76 et 1975-76 à 1980-81) et une période de onze années (1970-71 à 1980-81). On calcule les taux de changement de longue durée pour chaque période en comparant l'état des entreprises pendant les années initiale et finale. Ainsi, pour la période 1970-71 à 1980-81, le taux d'entrée est calculé comme le total des effectifs en 1981 dans les entreprises manufacturières qui ne faisaient pas partie du secteur manufacturier en 1970, divisés par le total des effectifs en 1970 dans ce secteur. Le résultat prend en compte l'effet cumulatif de tous les entrants entre 1971 et 1981 qui existaient encore en 1981.

La Figure 3 correspond à un graphique à barre illustrant la contribution cumulative totale au chapitre de l'emploi selon l'entrée des entreprises et l'expansion des entreprises en exploitation, ainsi que selon la contraction des entreprises en exploitation et la sortie des entreprises entre les années 1970 et 1981. Les entrées, en 1970, ont augmenté les effectifs de 10.9 pour cent et l'expansion, de 27.2 pour cent: les sorties ont entraîné la perte de 10.5 pour cent des emplois, et la contraction, de 11.0 pour cent des niveaux d'emplois initiaux.

Les résultats indiqués dans ce document ne rendent compte que d'une partie de la recherche faite à l'aide des données longitudinales du recensement. Ils permettent néanmoins d'illustrer une conclusion qui ressort du travail effectué. Bien que l'activité des entreprises qui sont encore en exploitation soit d'une grande importance, plus la période choisie est longue, plus le processus d'entrée et de sortie dépend de l'activité des entreprises en exploitation. L'importance de l'entrée et de la sortie ne se manifeste que sur une longue période.

L'entrée et la sortie v. l'expansion et la contraction des entreprises en exploitation: longue durée (effet cumulatif pour la période de 1970-81)

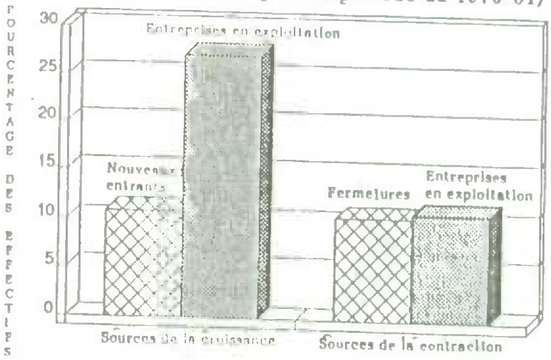


Figure 3

Il est donc important de pouvoir créer des bases de données longitudinales qui suivent l'activité des usines et des entreprises sur une période donnée. Cela n'est pas facile à faire. Si l'on doit évaluer les résultats de ces exercices, il faut examiner complètement le mode de création de ces bases. Nous traiterons dans le reste de ce document de la méthode qui a été utilisée pour constituer les bases de données qui ont servi à mesurer l'entrée et la sortie en particulier, et la rotation des entreprises en général (reportez-vous aux documents de recherche de Baldwin et Gorecki, 1989a à 1989g).

Taux de changement de courte durée

Afin de représenter un changement de courte durée, les taux d'entrée et de sortie sont calculés annuellement de 1970 à 1982. Les taux d'expansion et de contraction des entreprises en exploitation sont également calculés pour la même période.

Nous mesurons la dynamique intra-industrielle en nous concentrant sur les changements dans l'emploi pour chaque catégorie. Les changements dans l'emploi sont mesurés au niveau de l'entreprise consolidée. Un nouvel entrant correspond à une entreprise qui entre pour la première fois dans le secteur manufacturier en construisant une usine. Un sortant pour cause de fermeture équivaut à une entreprise qui quitte définitivement le secteur manufacturier en fermant une usine.

Les taux sont calculés comme le pourcentage des effectifs totaux du secteur provenant des entrants, des sortants, de la croissance des entreprises qui connaissent une expansion ou du déclin des entreprises qui subissent une contraction. La Figure 1 nous permet de comparer le taux moyen annuel d'expansion des nouveaux entrants à celui des entreprises en expansion et le taux moyen annuel de contraction des sortants pour cause de fermeture à celui des entreprises qui périclitent.

En moyenne, au cours des années 1970, l'entrée rend compte de 0.9 pour cent de l'emploi annuel, l'expansion des entreprises en exploitation, de 7.8 pour cent; la sortie correspond à 1.1 pour cent et la régression, à 6.3 pour cent de l'emploi annuel.

Ces chiffres démontrent clairement que les taux annuels d'entrée ne sont pas suffisamment importants pour indiquer qu'un changement même modéré peut être attribuable aux entrants.

Développement du processus d'entrée

Les petites valeurs que l'on obtient pour les taux d'entrée et de sortie instantanés ou de courte durée ne sont pas surprenantes. Elles viennent confirmer l'impression fortuite que les entrants parviennent rarement à dominer une industrie pendant leur première année d'exploitation. On pourrait s'en servir pour appuyer l'idée que l'entrée n'est pas un facteur important, mais cela ne serait pas justifié à ce moment-ci. Une telle conclusion doit se fonder sur d'autres éléments que le taux d'entrée instantané. Il faut également voir si ces nouvelles entreprises réussissent à croître sur une plus longue période et à supplanter des entreprises existantes, et déterminer le temps qu'elles ont pris pour le faire.

On peut calculer les mesures de longue durée de l'entrée à partir de la part du marché accumulée par tous les entrants depuis une année initiale. La part totale des entrants augmentera sur une période donnée parce que d'autres cohortes viennent s'ajouter chaque année; toutefois, cette tendance peut être neutralisée si la part des cohortes existantes du marché baisse. Si chaque cohorte augmente l'emploi de n pour cent, en moyenne, à compter de la période zéro puis que ce pourcentage baisse constamment de m points par année, la valeur cumulative maximale que l'entrée peut enregistrer se situe dans la n/m^{ième} période.

L'entrée et la sortie v. l'expansion et la contraction des entreprises en exploitation: courte durée (moyenne annuelle pour la période 1970-82)

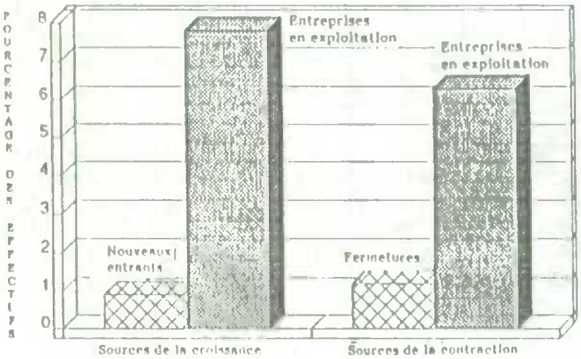


Figure 1

La part d'une cohorte particulière d'entrants sur une longue durée sera fonction du taux de sortie (fermetures), de la durée de vie moyenne et du taux de croissance que l'on peut constater après la création de tous les entrants dans cette cohorte. Si les entrants ont une durée de vie relativement courte en raison du taux élevé de fermetures aussitôt après leur création ou s'ils connaissent un taux de croissance relativement lent au cours des premières années, l'effet cumulatif ou de longue durée de l'entrée peut être négligeable. D'autre part, les entrants qui restent en exploitation peuvent se développer suffisamment pour contrebalancer l'effet des sorties et permettre à la part occupée par une cohorte d'augmenter sur une période assez longue. Dans ce cas, l'effet cumulatif de l'entrée sera plus grand.

Afin de pouvoir caractériser l'expérience des entrants qui sont restés en exploitation au cours des années 1970, nous avons utilisé les données d'entrée dans le secteur manufacturier en général et de sortie de ce secteur pour calculer la part de chaque cohorte d'entrants au fur et à mesure de son développement. Nous nous sommes servis des données de chaque cohorte d'entrants entre 1971 et 1980 et avons calculé la part moyenne, en termes de nombre d'entreprises et de valeur ajoutée, pour chaque groupe d'années d'exploitation de chaque cohorte d'entrants. Etant donné qu'il y a une sortie immédiate de chaque cohorte d'entrants, le pourcentage moyen de toutes les entreprises prises en compte par chacune d'elles diminue continuellement avec le nombre d'années de

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MESURE DE L'ENTRÉE ET DE LA SORTIE D'ENTREPRISES À L'AIDE DE DONNÉES LONGITUDINALES

J.R. Baldwin et P.K. Gorecki[1]

RÉSUMÉ

La dynamique du processus concurrentiel peut plus facilement se comprendre si l'on étudie l'entrée et la sortie, la croissance et le déclin des entreprises existantes, l'effet des fusions et l'importance du processus de rotation par rapport à l'accroissement de la productivité. De telles études doivent faire appel à des bases de données longitudinales qui mesurent la performance des entreprises sur une période donnée. Le présent document décrit la méthode utilisée pour créer ce genre de base de données à partir de renseignements provenant du recensement des manufactures. La construction d'un panel longitudinal à partir de données qui n'ont pas été recueillies à cette fin n'est pas une tâche facile. Ce document trace les grandes lignes des difficultés que cela pose et des choix qui ont été faits pour résoudre celles-ci. Au fur et à mesure que d'autres études de la dynamique du processus concurrentiel sont réalisées au Canada et à l'étranger, de plus en plus de comparaisons peuvent être faites d'un pays à un autre. Le présent document a été établi dans le but de fournir au lecteur des études connexes qui portent sur le Canada, les moyens d'évaluer ces études et de comparer, le cas échéant, les résultats de celles-ci avec ceux d'études réalisées dans d'autres pays à l'aide d'autres sources de données.

MOTS CLÉS: Entrée d'entreprises; sortie d'entreprises; création de bases de données.

IMPORTANCE DE L'ENTRÉE ET DE LA SORTIE

INTRODUCTION

On a longtemps cru que le processus d'entrée et de sortie des entreprises et des usines jouait un rôle important dans l'évolution de l'industrie et l'adaptation de celle-ci aux changements. Dans sa version la plus simple, la mesure de l'entrée et de la sortie permet de mettre en équation les profits supérieurs et inférieurs à la normale avec les taux concurrentiels. Dans d'autres modèles, l'entrée potentielle plutôt que réelle permet de limiter le pouvoir des monopoles. Une fois ce raisonnement placé sous la rubrique des modèles de fixation du prix limite, la théorie de la disputabilité lui a donné une élégance théorique. Le processus de rotation qui est le résultat de l'entrée et de la sortie est également perçu comme un conduit par lequel arrivent de nouvelles idées et des innovations.

De la même façon, on peut décrire l'entrée comme une curiosité intéressante, mais non pertinente. Une telle description représente les entrants comme des entreprises marginales qui envahissent l'industrie et s'en retirent sans laisser trop de traces. Comparer le processus d'entrée et de sortie à un "délit de fuite" donne une impression, voulue ou non, d'instabilité qui ne modifie en rien les indicateurs de progrès, telle que la productivité. Shepherd (1984), dans une critique de la théorie de la disputabilité, fait ressortir que l'entrée considérée comme une force externe est habituellement un facteur secondaire par rapport aux conditions internes qui prévalent dans l'industrie quand il s'agit de déterminer la force de la concurrence au sein d'une industrie.

En dépit de l'importance potentielle du processus d'entrée, ce n'est que tout récemment que celui-ci a attiré un peu d'attention sur le côté empirique de la littérature de l'organisation industrielle. Cette attention, toute neuve, reflète un plus grand intérêt de la part des économistes industriels pour la dynamique des marchés (comment les entreprises et les industries se comportent sur une période donnée et quels en sont les effets sur la structure et le comportement de l'industrie).

Compte tenu de la pénurie de données empiriques relatives au processus d'entrée, le débat sur l'importance de l'entrée n'est toujours pas éclairci. Un des problèmes que pose l'évaluation de l'importance de l'entrée et de la sortie est attribuable au manque de données longitudinales qui jalonnent la vie des entreprises dans le temps. Le recencement canadien des manufactures et ses équivalents dans d'autres pays, sont conçus pour recueillir et présenter des données agrégées de l'industrie à un moment donné et ne pouvaient pas, jusqu'à tout récemment, suivre les changements de chaque micro-unité sur une période donnée. Heureusement, le recensement canadien et ses dossiers connexes renferment des identificateurs pour chaque établissement et entreprise qui offrent la possibilité de créer un panel longitudinal. Nous faisons, dans ce document, un bref compte rendu de quelques-uns des résultats d'un projet où l'on s'est servi de ces données pour mesurer l'importance de l'entrée et de la sortie, en nous attachant principalement à décrire comment les bases de données ont été créées. Le fait de pouvoir compter sur la présence d'identificateurs ne permet pas en soi de réaliser des études longitudinales, surtout si les identificateurs n'ont pas été créés à cette fin. De telles bases de données servent trop souvent à faire des travaux de recherche sans être bien documentées.

[1] Groupe de l'Analyse des entreprises et du marché du travail. J.R. Baldwin est professeur d'économie à la Queen's University, Kingston, (Ontario), Canada K7L 3N6; P.K. Gorecki occupe un poste d'économiste supérieur au Conseil économique du Canada, Ottawa, (Ontario), Canada K1P 5V6.

[1] Selon le comité de la méthodologie statistique, l'objet fondamental de l'enquête longitudinale est, dès le départ, de déterminer des valeurs futures pour chaque unité d'observation (1986). Le comité a comparé des enquêtes longitudinales avec des enquêtes qui servent de base à l'analyse longitudinale. La base de données longitudinales (LRD) fait partie de la seconde catégorie. Parmi les 12 séries de données étudiées, elle est la seule qui porte sur des établissements.

[2] La plupart des études qui utilisent des données de panel reposent sur des données et des modèles qui se rapportent à des personnes. Bien qu'un grand nombre des méthodes qui s'appliquent à des personnes puissent être étendues à des modèles de comportement des entreprises et des établissements, de nouvelles questions surgissent. Par exemple, la relation étroite entre le changement de propriétaire et le comportement de l'établissement n'a pas d'équivalent du côté des personnes. Dans la plupart des cas, on pourra considérer le ménage comme unité d'analyse. Toutefois, l'analogie ne sera pas parfaite.

[3] Les statisticiens préconisent aussi l'utilisation des données de panel pour réduire la colinéarité et accroître la précision des estimations dans les modèles économiques dynamiques qui comprennent des variables explicatives retardées.

[4] Les panels de 1984 et de 1989 contiennent avec une probabilité égale à un les 500 plus grandes entreprises enregistrées comme telles en 1984 et en 1989.

[5] Même si des modifications sont apportées chaque année à l'échantillon pour tenir compte de la création ou de la fermeture d'établissements, il y a toujours un certain retard dans la mise à jour et il faut parfois attendre le recensement suivant avant de régulariser la situation.

[6] Ce nombre représente environ 30% de l'échantillon d'établissements de l'EAM.

[7] Malgré cela, le fait d'insister depuis toujours sur les totalisations agrégées a un effet défavorable sur les couplages disponibles.

[8] Les données des rapports industriels courants (Current Industrial Reports) ne sont pas raccordées à la LRD. Ces rapports contiennent des données annuelles et parfois mensuelles (valeur unitaire) pour de nombreux sous-groupes de la CAE. Le CES entend se servir de ces données dans le cadre de plusieurs projets précis et espère les raccorder éventuellement à la LRD.

[9] Des études récentes permettent de croire que les prix varient selon les établissements et les régions. L'établissement pourrait donc être le niveau de référence tout indiqué pour certains projets de recherche (voir Abbott (1989)).

[10] Les études économiques de ce genre sont rares. On les retrouve surtout dans le domaine de la finance. Voir McGuckin, Warren- Boulton et Waldstein (1988) pour un exemple d'étude de cas où l'on utilise des données de la bourse des valeurs mobilières.

[11] Pour diverses raisons ayant trait au plan d'échantillonnage de l'EAM, les comparaisons fondées sur un recensement et les passages de l'EAM qui le précèdent auront pour effet de gonfler le taux de transfert annuel. Malgré tout, ce chiffre est élevé.

[12] Ces chiffres sont établis à l'aide des données relatives aux quelque 50,000 établissements échantillonnés dans le panel de l'EAM. Les totaux relatifs à la production pour 1981 sont les totaux publiés par le BEA pour les catégories de produits. Les 1337 catégories qui ont servi à cette étude comprenaient toutes les catégories pour lesquelles il existait des données complètes pour chaque année et qui avaient la même définition d'une année à l'autre. Environ 200 catégories ont été éliminées à l'étape de la vérification.

[13] Dans les études de cas de fusion mentionnées plus haut, beaucoup de transferts d'une industrie à une autre ou d'une catégorie de produit à une autre sont le résultat d'un changement de propriété.

[14] De plus, le plan de sondage de l'EAM contient des règles qui restreignent la reclassification d'établissements dans les années intercensitaires. Par ailleurs, il est très difficile de constater l'existence de nouveaux établissements avant que l'enquête sur l'organisation des établissements n'ait été réalisée.

[15] L'utilisation d'un quotient fondé sur quatre entreprises plutôt que trois ou deux est dictée par les règles de protection du secret statistique appliquées par le U.S. Bureau of the Census.

[16] Leur étude englobe tous les producteurs et non seulement les producteurs primaires. En revanche, elle ne tient pas compte des entreprises les moins importantes, c'est-à-dire celles qui, globalement, représentent moins de 1% de la production totale.

[17] Bien que nous n'en faisions pas mention dans cet article, des études ont été réalisées au Canada à l'aide d'une base de données semblable à la LRD. Ces études donnent à penser que les mesures brutes de flux sont extrêmement utiles pour l'analyse de la concurrence, de la productivité des exportations et du reclassement de la main-d'oeuvre. Des études de Baldwin et Gorecki (1989b, 89c, 89d, 89e) donnent à croire que le Canada et les États-Unis ont connu la même expérience dans les années 1970.

Tableau 5: Pourcentage de la production d'une catégorie de produit attribué à des usines qui ont, dans un sens ou dans l'autre, changé de catégorie de produit entre 1981 et 1982 - Usines appariées

```
                                                              Fréquence de la variation nette
  Valeur absolue du transfert net de production                 Effectif                Pourcentage
     - centre de classe (en pourcentage)            Effectif     cumulé    Pourcentage    cumulé

    0 | ****************************************       792         792       59.24         59.24
    1 | ******                                          109         901        8.15         67.39
    2 | ****                                             80         981        5.98         73.37
    3 | ***                                              58       1,039        4.34         77.71
    4 | **                                               49       1,088        3.66         81.38
    5 | ****                                             84       1,172        6.28         87.66
   10 | ***                                              65       1,237        4.86         92.52
   15 | *                                                29       1,266        2.17         94.69
   20 | *                                                13       1,279        0.97         95.66
   25 |                                                   8       1,287        0.60         96.26
   30 | *                                                15       1,302        1.12         97.38
   35 |                                                   5       1,307        0.37         97.76
   40 |                                                   4       1,311        0.30         98.06
   45 |                                                   4       1,315        0.30         98.35
   50 |                                                   4       1,319        0.30         98.65
   55 |                                                   0       1,319        0.00         98.65
   60 |                                                   1       1,320        0.07         98.73
   65 |                                                   2       1,322        0.15         98.88
   70 |                                                   1       1,323        0.07         98.95
   75 |                                                   0       1,323        0.00         98.95
   80 |                                                   2       1,325        0.15         99.10
   85 |                                                   1       1,326        0.07         99.18
   90 |                                                   0       1,326        0.00         99.18
   95 |                                                   0       1,326        0.00         99.18
  100 | *                                                11       1,337        0.82        100.00
        -----+-----+-----+-----+-----+-----+-----+-----+
        100   200   300   400   500   600   700   800
                       Effectif
```

* On calcule le transfert net de production pour une catégorie de produit donnée en exprimant la différence entre la production de 1981 des établissements qui ne produisent plus la même catégorie de produit en 1982 et la production de 1982 des établissements qui produisaient d'autres catégories de produits en 1981 en pourcentage de la production de 1981 pour cette catégorie de produit.

Tableau 6: Taux de survie des entreprises les plus importantes, années diverses

Entreprises classées parmi les plus importantes en 1977 et qui étaient toujours en exploitation en 1982

Degré de concentration	20 plus grandes entreprises		8 plus grandes entreprises		4 plus grandes entreprises	
de l'industrie en 1977	Nbre	%	Nbre	%	Nbre	%
Supérieur à .6	11.7	58.5	6.0	75.0	3.3	82.5
Entre .4 et .6	11.3	56.5	6.0	75.0	3.2	80.0
Inférieur à .4	9.7	48.5	5.6	70.0	3.0	75.0
Total	11.4	57.0	5.8	72.5	3.1	77.5

Entreprises classées parmi les plus importantes en 1972 et qui étaient toujours en exploitation en 1977

Degré de concentration	20 plus grandes entreprises		8 plus grandes entreprises		4 plus grandes entreprises	
de l'industrie en 1972	Nbre	%	Nbre	%	Nbre	%
Supérieur à .6	12.2	61.0	6.4	80.0	3.5	87.5
Entre. 4 et .6	11.7	58.5	6.5	81.3	3.5	87.5
Inférieur à .4	10.4	52.0	5.7	71.3	3.2	80.0
Total	11.8	59.0	6.2	77.5	3.3	82.5

Entreprises classées parmi les plus importantes en 1972 et qui étaient toujours en exploitation en 1982

Degré de concentration	20 plus grandes entreprises		8 plus grandes entreprises		4 plus grandes entreprises	
de l'industrie en 1972	Nbre	%	Nbre	%	Nbre	%
Supérieur à .6	8.8	44.0	5.1	63.8	2.9	72.5
Entre .4 et .6	8.0	40.1	5.1	63.8	2.9	72.5
Inférieur à .4	6.6	33.0	4.1	51.3	2.5	62.5
Total	8.4	42.0	4.7	58.8	2.7	67.5

Tableau 3: Pourcentage de la production d'une catégorie de produit pour 1981 attribué aux usines qui ont changé de catégorie de produit en 1982 - Usines appariées*

Valeur de la production transférée - centre de classe (en pourcentage)		Fréquence de la variation brute			
		Effectif	Effectif cumulé	Pourcentage	Pourcentage cumulé
0	***************	110	110	8.23	8.23
1	****************	119	229	8.90	17.13
2	***************	110	339	8.23	25.36
3	*************	99	438	7.40	32.76
4	**************	102	540	7.63	40.39
5	****************************	212	752	15.86	56.25
10	*****************************	220	972	16.45	72.70
15	*****************	126	1,098	9.42	82.12
20	***********	83	1,188	6.21	88.33
25	******	47	1,228	3.52	91.85
30	*****	36	1,264	2.69	94.54
35	*	11	1,275	0.82	95.38
40	**	16	1,291	1.20	96.56
45	**	14	1,305	1.05	97.61
50		3	1,308	0.22	97.83
55	*	6	1,314	0.45	98.28
60	*	5	1,319	0.37	98.65
65	*	4	1,323	0.30	98.95
70	*	5	1,328	0.37	99.33
75		1	1,329	0.07	99.40
80		1	1,330	0.07	99.48
85		2	1,332	0.15	99.63
90		1	1,333	0.07	99.70
95		0	1,333	0.00	99.70
100	*	4	1,337	0.30	100.00
	--------+---------+---------+---------				
	60 120 180 Effectif				

* On calcule le transfert brut de production en exprimant la production de 1981 des usines qui ne produisent plus la même catégorie de produits en 1982 en pourcentage de la production totale de 1981 pour cette catégorie de produits.

Tableau 4: Pourcentage de la production d'une catégorie de produit pour 1982 attribué aux usines qui fabriquaient d'autres catégories de produits en 1981 - Usines appariées*

Valeur de la production transférée - centre de classe (en pourcentage)		Fréquence de la variation brute			
		Effectif	Effectif cumulé	Pourcentage	Pourcentage cumulé
0	***************	113	131	9.80	9.80
1	*******************	145	276	10.85	20.64
2	******************	135	411	10.10	30.74
3	*****************	127	538	9.50	40.24
4	*************	95	633	7.11	47.34
5	******************************	226	859	16.90	64.25
10	****************************	209	1,068	15.63	79.88
15	***********	85	1,153	6.36	86.24
20	*********	64	1,217	4.79	91.02
25	*****	37	1,254	2.77	93.79
30	**	17	1,271	1.27	95.06
35	**	17	1,288	1.27	96.34
40	*	6	1,294	0.45	96.78
45	*	8	1,302	0.60	97.38
50	*	7	1,309	0.52	97.91
55		2	1,311	0.15	98.06
60		3	1,314	0.22	98.28
65		2	1,316	0.15	98.43
70	*	2	1,318	0.15	98.58
75		1	1,319	0.07	98.65
80		0	1,319	0.00	98.65
85		2	1,321	0.15	98.80
90		0	1,321	0.00	98.80
95		3	1,324	0.22	99.03
100	**	13	1,337	0.97	100.00
	--------+---------+---------+---------				
	60 120 180 Effectif				

* On calcule le transfert brut de production en exprimant la production de 1982 des établissements qui produisaient d'autres catégories de produits en 1981 en pourcentage de la production totale de 1981 pour cette catégorie de produits.

Sigles	Variables	Données disponibles*
ph3	heures-personnes: juillet-septembre	
ph4	heures-personnes: octobre-décembre	
ph	nombre total d'heures-personnes	
sw	rémunération totale	
ww	salaires: travailleurs affectés à la production	
ow	salaires: autres employés	
lc	revenu supplémentaire du travail total	
le	revenu supplémentaire du travail exigé par la loi	
vlc	revenu supplémentaire du travail gagné volontairement	
cp	coût des matières, pièces, etc.	
cr	coût des marchandises revendues	
cf	coût du combustible	
ee	coût de l'électricité achetée	
pe	quantité d'électricité achetée	
cw	coût lié à l'exécution de contrats	
cpc	coût des communications	A 77 et 82
fib	stock au début: produits finis	
wib	produits en cours	
mib	matières	
fie	stock à la fin: produits finis	
wie	produits en cours	
mie	matières	
tib	stock au début: total	
tie	stock à la fin: total	
nb	dépenses en immobilisations (construction de bâtiments)	
nm	dépenses en immobilisations (achat de matériel)	
ue	dépenses en immobilisations (rénovation et remise en état)	
bab	actif immobilisé (bâtiments) - début de l'exercice	A; après 1973
mab	actif immobilisé (matériel) - début de l'exercice	A; après 1973
bae	actif immobilisé (bâtiments) - fin de l'exercice	A
mae	actif immobilisé (matériel) - fin de l'exercice	A
br	loyer - bâtiments	A
mr	loyer - matériel	A
bd	amortissement - bâtiments	A; après 1976
md	amortissement - matériel	A; après 1976
brt	mise hors service - bâtiments	A; après 1976
mrt	mise hors service - matériel	A; après 1976
rbs	réparations - bâtiments	A; après 1976
rm	réparations - matériel	A; après 1976
m	matières	C
mqpc	quantité produite et utilisée	C
mqdc	quantité reçue et utilisée	C
mc	coût à la livraison	C
pi	code du produit	C
pqp	quantité produite	C
pqs	quantité expédiée	C
pv	valeur expédiée	C
pgit	quantité transférée entre usines	C
pvit	valeur des produits transférés entre usines	C
pqpc	quantité produite et utilisée	C
tvs	valeur totale des expéditions	C

* Des données existent pour toutes les années et tous les établissements sauf dans le cas où ces lettres sont indiquées:
A = données recueillies uniquement pour les établissements visés par l'EAM;
C = données recueillies uniquement dans les années de recensement.

McGuckin, Robert H., Warren-Boulton, Frederick R., and Waldstein, Peter (1988). "Analysis of Mergers Using Stock Market Returns," Economic Analysis Group Discussion Paper, EAG 88-1.

Nguyen, Sang V. and Reznek, Arnold P. (1989), "Production Technologies, Economies of Scale, and Factor Substitution in Large and Small U.S. Manufacturing Establishments: A Pilot Study," Forthcoming in Center for Economic Studies Discussion Paper.

Tableau 1: Nombre d'établissements contenus dans la LRD pour chaque année

Année	Nombre d'établissements	Nombre de cas fondés sur des dossiers administratifs
1963	305,747	*
1967	305,611	118,622
1972	312,398	122,158
1973	73,460	-
1974	68,262	-
1975	71,145	-
1976	70,346	-
1977	350,648	144,648
1978	73,853	-
1979	57,559	-
1980	55,953	-
1981	55,045	-
1982	348,384	128,307
1983	51,619	-
1984	56,551	-
1985	55,128	-
1986	54,858	-

* Aucun cas de ce genre en 1963

- L'EAM ne comprend pas de cas fondés sur des dossiers administratifs.

Tableau 2: Variables contenues dans la LRD

Sigles	Variables	Données disponibles*
ppn	code permanent de l'usine	
id	code d'identification	
ind	code de l'industrie	
ppc	catégorie de produit primaire	
pisr	ratio de spécialisation - industrie primaire	
ppsr	ratio de spécialisation - produit primaire	
ii3	condition de l'établissement	
ei	code d'identification de l'employeur	
dind	code de l'industrie dérivé	
et	genre d'établissement (0=EAM)	C
ar	dossier administratif (1=DA)	C
cc	code de couverture	
sc	code de source	
lfo	forme juridique de l'organisation	C
st	code de l'État	
smsa	code de la smsa	
cou	code du comté	
plac	code de la localité	
va	valeur ajoutée	
vr	valeur de revente	
rcw	recettes tirées de l'exécution de contrats	
msc	recettes diverses	
te	emploi total	
pw1	travailleurs affectés à la production: mars	
pw2	travailleurs affectés à la production: mai	
pw3	travailleurs affectés à la production: août	
pw4	travailleurs affectés à la production: novembre	
pw	travailleurs affectés à la production (moyenne)	
ph1	heures-personnes: janvier-mars	
ph2	heures-personnes: avril-juin	

Blair, Margaret (1989). "Free Cash Flow and the Rise in Contests for Corporate Control," Brookings Discussion Paper in Economics.

Baldwin, John R. and Gorecki, Paul K. (1987). "The Dynamics of Firm Turnover," Economic Council of Canada Working Paper.

Baldwin, John R. and Gorecki, Paul K. (1989a). "Measures of Market Dynamics: Concentration and Mobility Statistics for the Canadian Manufacturing Sector," Forthcoming *Annales D' Economic et Statistique.*

Baldwin, John R. and Gorecki, Paul K. (1989b). "Job Turnover in Canada's Manufacturing Sectors," Statistics Canada Analytical Studies Branch Research Paper Series, No. 22.

Baldwin, John R. and Gorecki, Paul K. (1989c). "Firm Entry and Exit in the Canadian Manufacturing Sector," Statistics Canada Analytical Studies Branch Research Paper Series, No. 23.

Baldwin, John R. and Gorecki, Paul K. (1989d). "Mergers Placed in the Context of Firm Turnover," Department of Economics, Queen's University, and Statistics Canada, Draft.

Baldwin, John R. and Gorecki, Paul K. (1989e). "Dimensions of Labor Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover," Business and Labor Market Analysis Group Statistics Canada, Draft.

Caves, R.E. and Porter, M.E. (1978). "Market Structure, Oligopoly, and Stability of Market Shares," *Journal of Industrial Economics*, Volume XXVI, pp 289-313.

Caves, Richard E. and Porter, Michael E. (1980). "The Dynamics of Changing Seller Concentration," *Journal of Industrial Economics*, Volume XXIX, pp 1-15.

Davis, Steve J. and Haltiwanger, John (1989). "Gross Job Creation, Gross Job Destruction and Employment Reallocation," Forthcoming in Center for Economic Studies Discussion Paper.

Dunne, Timothy and Roberts, Mark J. (1986). "Measuring Firm Entry and Exit With Census of Manufacturers Data," Department of Economics, The Pennsylvania State University.

Dunne, Timothy and Roberts, Mark J. (1987). "The Duration of Employment Opportunities In U.S. Manufacturing," Department of Economics, The Pennsylvania State University.

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1988a). "The Growth and Failure of U.S. Manufacturing Plants," Department of Economics, The Pennsylvania State University Working Paper.

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1988b). "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries," *The Rand Journal of Economics*, Vol. 19, No. 4 (WINTER).

Dunne, Timothy, Roberts, Mark J. and Samuelson, Larry (1989). "Firm Entry and Post-Entry Performance in the U.S. Chemical Industries," Center for Economic Studies Discussion Paper, CES 89-6.

Gort, Michael (1973). "Analysis of Stability and Change in Market Shares," *Journal of Political Economy*, Volume 71, pp 51-63.

Gossack, Irvin M. (1965). "Towards an Integration of Static and Dynamic Measures of Industry Concentration," *Review Economics and Statistics*, Volume 47, pp 301-308.

Lichtenberg, Frank and Siegel, Donald (1987). "Productivity and Changes in Ownership of Manufacturing Plants," *Brookings Papers on Economic Activity*, pp 643-673.

Lichtenberg, Frank R. and Siegel, Donald (1989a). "The Effect of Takeovers on the Employment and Wages of Central-Office and Other Personnel," Center for Economic Studies Discussion Paper, CES 89-3.

Lichtenberg, Frank R. and Siegel, Donald (1989b). "The Effects of Leveraged Buyouts on Productivity and Related Aspects of Firm Behavior," Center for Economic Studies Discussion Paper, CES 89-5.

McGuckin, Robert H. (1972). "Entry, Concentration Change, and Stability of Market Shares," *Southern Economic Journal*, Vol. XXXVIII, No. 3.

McGuckin, Robert H. and Nguyen, Sang V. (1988). "Public Use Microdata: Disclosure and Usefulness," Center for Economic Studies Discussion Paper, CES 88-3.

McGuckin, Robert H. and Andrews, Stephen H. (1988). "The Performance of Lines of Business Purchased in Conglomerate Acquisitions," paper presented at the American Economic Association Meeting in Chicago.

McGuckin, Robert H. and Pascoe, George A. Jr. (1988). "The Longitudinal Research Data Base (LRD): Status and Research Possibilities," *Survey of Current Business*, Vol. 68, pp 30-37.

taux de survie diffèrent peu entre eux selon le degré de concentration de l'industrie. Bien que le taux de survie le moins élevé corresponde toujours au cas où le degré de concentration est inférieur à 0.4, il y a peu de différence entre les taux de survie qui correspondent aux deux autres cas et l'écart entre ceux-ci et le taux correspondant au troisième cas n'est pas non plus très prononcé. De plus, peu importe le degré de concentration initial dans l'industrie, l'industrie moyenne a vu disparaître au moins une de ses quatre plus grandes entreprises au début de chaque période.

Bien que l'on ne puisse pas faire de comparaison directe à cause des différences de méthode, mentionnons que Dunne, Roberts et Samuelson (1988) ont eux aussi défini des taux d'entrée et de sortie bruts pour les industries du niveau d'agrégation à 4 chiffres[16]. Leur méthode se distingue principalement de la nôtre par la façon de considérer les changements de propriété. Dunne et coll. ne considèrent pas les entreprises qui changent de mains comme de nouvelles venues, à moins que le changement de propriété ne modifie la structure fondamentale des entreprises sur le marché. Les seules entreprises qu'ils considèrent comme de nouvelles venues sont celles qui contribuent à accroître la capacité sur le marché. Lorsque la direction d'une entreprise change, ils considèrent cela comme un changement de nom.

Par contre, dans le présent article,'toutes les entreprises qui changent de mains sont classées parmi les entrants ou les sortants pour le calcul des taux de survie. Si l'on entend par nouveau compétiteur toute entreprise qui contribue à l'accroissement de la capacité de production, il est donc logique de ne pas tenir compte des "changements de nom" qui résultent d'une fusion ou de toute autre forme de changement de propriété. Toutefois, si, comme le laisse à entendre l'étude sur les changements de propriété citée plus haut, le changement de propriété amène de nouveaux dirigeants et un meilleur rendement, les "changements de nom" devraient être assimilés aux entreprises nouvelles venues.

Mais dans quelle mesure les assimiler est une tout autre question, qui ne peut être réglée dans l'état actuel des choses. Ce n'est qu'après d'autres études empiriques mettant en relation le rendement et les mesures de comportement d'une part, et les taux de survie et les autres indices de concentration dynamique d'autre part, que nous pourrons vraiment nous prononcer.

Une chose est sûre: les données de panel sont indispensables pour faire avancer la recherche dans ce domaine. Même si l'exemple que nous avons présenté porte essentiellement sur la variation transversale (ou variation d'une industrie à l'autre) du taux de rotation, comme l'étude de Davis et Haltiwanger (1989) citée plus haut, les variations des séries chronologiques, qui reflètent les fluctuations de la demande, les progrès technologiques et la fluctuation des prix des intrants seront vraisemblablement des éléments majeurs dans le calcul des taux de rotation brut et net. Cela est la conséquence évidente des vagues de fusions qui sont associées notamment à des bouleversements dans l'industrie (voir Blair (1989)).

4. CONCLUSIONS

Comme le laisse à entendre la citation du début, les panels qui existent dans la LRD permettent de réaliser toute une gamme d'études longitudinales. Dans cet article, nous nous sommes arrêtés à deux types d'études que l'on peut réaliser avec des données de panel. Le premier type est l'étude de cas. Dans les exemples cités, nous montrons l'importance d'utiliser des effets qui varient dans le temps pour expliquer le comportement des entreprises et des établissements (récemment créés ou non). Diverses études du CES ont montré qu'un changement de propriété pouvait entraîner une modification notable du rendement de l'entreprise. Cela donne à penser qu'il est nécessaire d'inclure les changements de propriété dans les modèles qui expliquent le comportement des entreprises et des établissements. En outre, comme le nombre de fusions et de changements de propriété varie beaucoup d'une année à l'autre, ces changements peuvent avoir des effets appréciables sur les données chronologiques agrégées.

A cet égard, nous notons également le grand nombre de transferts d'établissements d'une industrie à une autre. Ces transferts peuvent créer des fluctuations dans les séries chronologiques de la production globale des industries puisque pour diverses raisons, les effets de ces transferts sont surtout observés dans les années de recensement. De plus, comme il y a des données qui prouvent que ces transferts sont le résultat d'un changement de propriété ou d'autres événements du même genre, leur effet sur les mesures agrégées de la production ne se fait pas sentir uniquement au point de vue du traitement ou du plan de sondage. Il s'agit plutôt d'un phénomène qu'il faut expliquer par un modèle. Compte tenu du nombre accru de fusions et d'acquisitions observées dans les années 1980, il faudrait à tout le moins évaluer l'incidence des transferts sur les données agrégées. Nous avons noté que des études en cours au CES arrivent à la conclusion que le taux de rotation brut de la main- d'oeuvre a des conséquences importantes pour l'analyse du marché du travail et des cycles économiques. De plus, nous avons illustré l'importance des flux bruts par une simple mesure "dynamique" de la structure du marché qui a permis le calcul de taux d'entrée et de sortie bruts grâce à la LRD.[17] Dans ce genre d'analyse, comme dans les études de cas, nous "obtenons ce qu'il nous faut".

BIBLIOGRAPHIE

Andrews, Stephen H. and Thomas A. Abbott III (1988). "An Examination of the Standard Industrial Classification of Manufacturing Activity Using the Longitudinal Research Data Base, "*Bureau of the Census Fourth Annual Research Conference Proceedings.*

3.3.1 Reclassement de la main-d'oeuvre

Une étude récente de Davis et Haltiwanger (1989) donne à penser que les mesures brutes de création d'emploi sont importantes pour l'étude des cycles économiques et d'autres questions macroéconomiques. Davis et Haltiwanger constatent que le taux de diminution de l'emploi dans le secteur manufacturier a été beaucoup moins élevé que le taux de reclassement bruts de la main-d'oeuvre (somme des taux de création et de suppression d'emplois) entre 1972 et 1986. La forte différence entre les taux de reclassement bruts et les variations nettes observées (environ 10 points de pourcentage en faveur des taux de reclassement) suppose l'existence de flux de main-d'oeuvre appréciables entre les établissements, lesquels flux échappent à notre attention à cause de l'analyse des variations nettes. En outre, Davis et Haltiwanger constatent que le taux de reclassement brut suit une forte tendance contracyclique, ce qui contraste avec la tendance procyclique du taux de reclassement net. Ce qu'il faut retenir de l'étude de Davis et Haltiwanger est que les mesures brutes de création et de suppression d'emplois sont importantes pour l'étude des cycles économiques et d'autres questions macroéconomiques. Nous allons voir maintenant que les mesures brutes de flux sont tout aussi importantes pour l'examen des questions microéconomiques.

3.3.2 Entrée et sortie des entreprises

On fait valoir depuis longtemps l'importance de la structure du marché dans la détermination du rendement. Jusqu'à récemment, on déterminait souvent l'existence d'un pouvoir monopolistique à l'aide de mesures de la structure du marché comme les quotients de concentration. Un quotient de concentration sert à mesurer la part de la production issue, par exemple, des quatre entreprises les plus importantes dans un marché.[15] Dans sa plus simple expression, la théorie dit que le quotient de concentration permet d'évaluer avec quelle facilité les entreprises les plus importantes d'un secteur peuvent coordonner leurs politiques de prix.

Le fait d'évaluer le pouvoir monopolistique uniquement par le quotient de concentration soulève de nombreux problèmes. Parmi les principaux, notons celui, reconnu depuis longtemps, qui fait valoir le rôle important de l'entrée (ou de l'entrée potentielle) comme obstacle ultime pour les entreprises qui pratiquent des prix non concurrentiels. Jusqu'à récemment, on disposait de peu d'information pour créer des mesures d'entrée autres que la variation nette du nombre d'entreprises.

Pour obtenir une mesure dynamique de la structure du marché qui repose sur l'entrée et la sortie brutes, on pourrait par exemple calculer le nombre de grandes entreprises dans un marché qui subsistent d'une période à une autre. La justification théorique de cette mesure est qu'elle fournit de l'information sur la rotation des entreprises concurrentes dans un marché. Cette mesure n'est pas nouvelle et les tests empiriques dont nous exposons les résultats ici ne donnent qu'une idée d'ensemble. Cependant, ils illustrent l'importance des considérations longitudinales dans l'analyse de la structure du marché.

Nous avons calculé des taux de survie pour les quelque 450 industries manufacturières du niveau d'agrégation à quatre chiffres pour les années 1972 à 1977, 1977 à 1982 et 1972 à 1982. En réalité, les calculs portaient sur les 20 entreprises les plus importantes (au point de vue de la valeur des expéditions) dans chaque industrie et pour chaque année de recensement. Par conséquent, le nombre d'entreprises ayant subsisté est simplement le complément de la rotation brute des entreprises pour la période étudiée. Autrement dit, pour calculer la rotation brute, il suffit de faire la différence entre le nombre total d'entreprises (20) et le nombre d'entreprises qui se sont maintenues parmi les 20 premières. Par définition, la rotation nette est égale à zéro.

En utilisant les 20 entreprises les plus importantes, nous réduisons le risque d'erreur de classification pour les petites entreprises. En règle générale, les 20 entreprises les plus importantes d'une industrie représentent la majeure partie de la production de cette industrie. Dans 280 des 450 industries observées, les 20 entreprises les plus importantes représentent à elles seules plus de 60% de la production. De fait, il n'y a que 55 industries où les 20 entreprises les plus importantes représentent moins de 40% de la production. Pour les trois années de recensement étudiées (1972, 1977 et 1982), la proportion de la production imputée aux 20 entreprises les plus importantes se situait en moyenne autour de 75%.

Les résultats de nos calculs indiquent un taux de rotation appréciable parmi les entreprises les plus importantes. Le tableau 6 montre qu'en 5 années à peine, l'industrie moyenne a vu disparaître 8 ou 9 de ses 20 plus grandes entreprises. Cela signifie un taux de rotation brut d'environ 40% (8/20) pour les périodes 1972-1977 et 1977-1982. Si nous examinons les chiffres pour la période de 1972 à 1982, nous voyons que le taux de rotation brut approche 60% en moyenne, c'est-à-dire que 11 ou 12 des 20 plus grandes entreprises de l'industrie type ont disparu en l'espace de 10 ans.

Ces taux de rotation élevés ne sont pas nécessairement le résultat d'un large exercice du pouvoir de monopole car, malgré ces taux, les parts de marché des entreprises les plus importantes peuvent être très stables. C'est ce que nous déduisons de la ventilation des taux de survie selon le degré de concentration, présentée dans le tableau 6.

Comme on pouvait s'y attendre, on obtient un taux de rotation plus élevé lorsqu'on le calcule pour les 20 plus grandes entreprises, plutôt que pour les 8 ou les 4 plus grandes. Par conséquent, si nous lisons chaque ligne du tableau 6 de la gauche vers la droite, nous remarquons que le taux de survie (en pourcentage) va en augmentant comme nous passons des 20 plus grandes entreprises aux 8 plus grandes, aux 4 plus grandes. En revanche, les

Les tableaux montrent que la distribution des variations brutes est très étendue. En moyenne, l'équivalent de plus de 10% de la production rattachée à une catégorie de produit est passée à une autre catégorie de produit. Pour environ les trois quarts des catégories de produits, la proportion de la production touchée par un transfert est supérieure à 5%. Pour 5% des catégories, cette proportion est supérieure à 70%.

Pour ce qui est de la variation nette, l'effet est beaucoup moins marqué (en moyenne, moins de 3% de la production totale de 1981 pour une catégorie de produit). Cependant, comme nous pouvons le constater dans le tableau 5, la variation nette est relativement élevée pour certaines catégories de produits. Pour plus de 10% des quelque 1350 catégories de produits, la variation nette est supérieure à 5%.[12]

Les données des tableaux 3 à 5 couvrent tous les groupes de produits du niveau d'agrégation à 5 chiffres qui peuvent servir à l'analyse. On pourrait faire valoir qu'un ensemble aussi varié donne une image exagérée de la situation. C'est exactement le cas du tableau 5. Une fois que l'on a mis à l'écart toutes les catégories de produits dont le code se termine par 0 ou 9 (catégories diverses), on constate que le pourcentage de la production totale de 1981 touchée par des transferts d'une catégorie à une autre varie de -3.5 à +3.5% pour plus de 90% des catégories de produits qui restent. Néanmoins, pour toutes les catégories qui restent, le pourcentage de la production totale touchée par des transferts oscillait en moyenne autour de 9% avec un écart type approximatif de 15. Ce phénomène n'est pas uniquement le fait de catégories de produits mal définies.

Ces observations donnent à penser que l'industrie (en tant que critère de classification) ne peut être considérée simplement comme un effet invariant dans le temps. De plus, les transferts de production ne sont pas uniquement le résultat d'erreurs de mesure aléatoires (ou non aléatoires) engendrées par les difficultés liées à la Classification des activités économiques. La recherche n'a pas encore permis d'établir si la probabilité d'un transfert de production est plus forte au moment d'un changement de propriété qu'à tout autre moment dans l'histoire d'un établissement. Certaines données (McGuckin et Andrews (1988)) suggèrent des conclusions en ce sens. Par ailleurs, les dirigeants d'entreprises prennent des décisions économiques qui visent à orienter la capacité de production d'une usine vers de nouvelles activités. Cela est vrai au moment d'un changement de propriété. Cela est aussi vrai dans les décisions au jour le jour par lesquelles on modifie la gamme de produits fabriqués par l'usine à cause notamment, d'un changement de la demande. Cela donne à penser que pour certains problèmes à tout le moins, les transferts de production doivent être considérés comme une variable endogène ou un phénomène expliqué.[13]

3.2 Implications pour les données chronologiques agrégées

Il est essentiel de reconnaître que les recherches décrites ci- dessus ont des conséquences pour l'analyse de données agrégées. La première conséquence a un rapport direct avec l'applicabilité de modèles comme celui décrit par l'équation (1). L'analyse de données agrégées a recours à des hypothèses sur la nature et l'homogénéité du comportement des agents économiques. Les résultats d'études fondées sur la LRD permettent de croire que les hypothèses classiques sur la "représentativité" des données agrégées pourraient ne pas être justes dans les circonstances.

La seconde conséquence a trait au caractère de la série d'observations. Pour diverses considérations liées au traitement, les variations de la production ou de l'emploi (dans une industrie) attribuables à des transferts d'une industrie à une autre sont enregistrées pour la plupart dans les années de recensement. Cela n'est pas le fruit du hasard; en effet, à l'occasion de chaque recensement, on réalise une enquête exhaustive sur l'organisation des établissements et cette enquête s'adresse à tous les établissements commerciaux. Les entreprises doivent décrire tous les produits qui sortent de leurs usines. Le U.S. Bureau of the Census se sert des nouvelles données pour reclasser les usines de sorte que celles-ci reçoivent le bon questionnaire pour le recensement. C'est pourquoi un grand nombre de transferts de production sont rapportés dans les années de recensement.[14]

A cause des réalités du traitement et de la forte proportion de la production d'une industrie susceptible de passer à une autre industrie, les données publiées sur la production globale, l'emploi ou d'autres variables axées sur l'établissement subiront quelques variations entre les années de recensement et les années intercensitaires. C'est à la suite de l'observation de ces variations que le BEA a mis sur pied son projet qui a permis de recueillir les données ayant servi à la construction des tableaux 3, 4 et 5. Si on parvient à déterminer l'origine de ces variations, on disposera de données pour accroître la qualité des séries chronologiques existantes. On pourra du même coup travailler à la reconstitution de séries chronologiques que l'on pourra comparer aux séries non redressées tirées des enquêtes et des recensements à caractère transversal. Ces travaux devraient nous fournir des renseignements précieux pour l'interprétation des modèles de séries chronologiques.

3.3 Flux bruts

La LRD permet de déterminer les flux bruts et les flux nets de variables économiques comme la création d'emplois et l'entrée d'entreprises dans un marché. Le besoin d'analyser des mesures de variation brute a été à l'origine d'un certain nombre d'études au CES. Ces études concluent toutes que le fait d'insister sur les mesures transversales agrégées de la variation nette peut soustraire à notre attention d'importants phénomènes économiques.

lesquelles des données de panel comme celles contenues dans la LRD sont indispensables. Mais ce qui est plus important encore, c'est de constater que ce genre d'analyse est indispensable pour bien comprendre les grands phénomènes économiques et porter des jugements éclairés sur les programmes.

3.1 Évolution de l'entreprise dans le temps

Pour des raisons de commodité, prenons tout d'abord un modèle simple pour expliquer le rendement ou l'évolution d'une entité économique telle une usine ou une entreprise. Par conséquent, supposons que le rendement de l'usine i au temps t, Y_{it}, peut être décrit par la relation

$$Y_{it} = \alpha + \mu_j + \lambda_t + \Sigma_s \beta_s X_{sit} + \varepsilon_{it} \tag{1}$$

où X_{sit} sont des variables exogènes, α représente l'effet fixe par usine, qui est commun à toutes les usines, μ_j désigne un effet fixe invariant dans le temps, comme la propriété, l'industrie ou l'emplacement, qui est commun à un groupe d'usines, λ_t est un effet fixe qui varie dans le temps et qui est le même pour toutes les usines, et ε_{it} est un terme d'erreur. Ce modèle élémentaire suffit pour décrire les questions qui nous intéressent.

Nous pouvons notamment nous demander ce que les effets fixes définis dans le modèle permettent de contrôler. Ce qui nous intéresse particulièrement ici, c'est de savoir quels effets invariants dans le temps peuvent être représentés par μ_j. Une réponse qui nous vient rapidement à l'esprit est le groupe de la CAE auquel appartient l'usine, ou encore la propriété de l'usine. Or, aucune de ces réponses n'est satisfaisante.

3.1.1 Changement de propriété

La fréquence et l'ampleur des fusions observées récemment montrent clairement que les entreprises ou les usines changent souvent de mains. Considérer la propriété comme un effet invariant dans le temps ne pose pas de problème si la situation de l'usine (ou de l'entreprise) demeure relativement la même après le changement de propriétaire. Cependant, des études fondées sur la LRD révèlent qu'un changement de propriétaire a des effets notables sur le rendement d'une usine ou d'une entreprise.

McGuckin et Andrews (1988) notent un accroissement de la part de marché pour les entreprises qui fusionnent par rapport à celles qui ne le font pas, surtout dans le cas des prises de contrôle à 100%. Lichtenberg et Siegel (1988, 1989b) notent une amélioration de la productivité à la suite d'un changement de propriétaire. De plus, ils peuvent expliquer en majeure partie ce gain de productivité par la réduction du nombre d'employés administratifs et la diminution des salaires de cette catégorie d'employés (Lichtenberg et Siegel (1989a)). Ce genre d'"études de cas" sont irréalisables si on ne dispose pas d'un panel d'observations sur les établissements ou les entreprises.[10]

3.1.2 Changement d'industrie primaire

Diverses études ont aussi montré que de nombreux établissements passent d'une industrie à une autre. Environ le tiers du panel composé de plus de 16,000 établissements observés sans interruption de 1972 à 1986 a changé d'industrie primaire (niveau d'agrégation à quatre chiffres). Le fait de considérer l'industrie comme un effet fixe peut donc représenter une entorse au modèle.

Le panel équilibré de la LRD renferme normalement un trop grand nombre de gros établissements. Les transferts d'une industrie à une autre ne sont donc pas le propre des petits établissements, qui comptent pour peu dans la production totale. Abbott et Andrews (1988) indiquent que les usines qui sont passées d'une industrie primaire à une autre entre deux recensements représentaient plus de 3% de la production totale dans les périodes 1972-1977 et 1977- 1982. Dans le cas de certains groupes (niveau à 2 chiffres), 10% de la production de l'industrie type du niveau d'agrégation à 4 chiffres était touchée par des transferts d'établissements. Bref, ces transferts ont une incidence appréciable sur les chiffres de la production totale de nombreuses industries.

3.1.3 Changement de catégorie de produit

Par surcroît, les transferts observés ne sont pas uniquement la conséquence d'erreurs de mesure commises lorsqu'une usine qui a une production diversifiée est classée dans une nouvelle industrie du fait que son "principal" produit a changé. Bon nombre de ces transferts impliquent un changement de catégorie de produit. En comparant les données d'établissements appariés pour l'EAM de 1981 et le recensement de 1982, nous avons constaté, sur la foi de données rassemblées à l'occasion d'une étude du Bureau of Economic Analysis (BEA) du Département du commerce des É.-U., que la production brute touchée par les changements de catégorie de produit représentait en moyenne plus de 10% de la production totale en 1981.[11]

Les tableaux 3 et 4 contiennent des données sur le pourcentage de la production totale d'une catégorie de produit attribué aux usines qui fabriquaient cette catégorie de produit en 1981 mais qui l'ont remplacée par d'autres l'année suivante. On appelle cela la part de production transférée vers une autre catégorie de produit. La part de production transférée d'une autre catégorie de produit est définie de la même façon et est aussi calculée au moyen de la production totale de 1981 pour la catégorie de produit en question.

L'existence d'un plan avec renouvellement et le fait que la plupart des établissements ne sont pas échantillonnés pour l'EAM n'ont, en théorie, aucun effet sur les estimations agrégées transversales. En revanche, il devient plus difficile de suivre l'évolution des établissements dans le temps et les établissements pour lesquels il existe des données continues pour chaque année deviennent moins nombreux. Ce plan limite considérablement le nombre de panels uniformes que l'on peut tirer annuellement de la LRD. Pour la période 1972-1986, on compte à peine plus de 16,000 établissements pour lesquels il existe des données pour chaque année de cette période dans la LRD; ce nombre représente moins de 5% des établissements recensés à une année ou à une autre.[6]

Les établissements qui ne sont pas échantillonnés pour l'EAM figurent uniquement dans le RM. Compte tenu de ce qu'il y a déjà eu 5 recensements avant celui de 1987 et que les résultats de celui- ci seront publiés sous peu, il y a de bonnes chances que nous puissions fonder notre recherche sur des panels équilibrés avec des observations aux 5 ans.[7] Roberts et Monahan (1986) donnent la composition des raccordements qui existent pour les années de recensement 1972, 1977 et 1982. Sur les quelque 600,000 enregistrements d'établissements consignés au cours de ces trois années, il y en a environ 133,000 (ou 22%) qui reviennent à chaque année. Dunne et Roberts (1986) ont étendu les données correspondantes à la période 1963-1982. Pour cette période, on dispose d'environ 66,000 établissements raccordés pour constituer un panel équilibré. Bien que l'effectif du panel sera moindre pour 1987 à cause de l'attrition, le nombre d'établissements observés durant toute la période 1963-1987 devrait se maintenir au-dessus de 50,000.

2.2 Données

La LRD renferme divers renseignements sur les établissements. Les données sont enregistrées annuellement, à l'exception de celles sur l'emploi et le nombre d'heures travaillées, qui le sont trimestriellement. De façon générale, les données contenues dans la LRD portent sur la production et diverses caractéristiques de classification et d'identification. Dans ce dernier cas, il peut s'agir de données sur la propriété de l'usine, son emplacement ou son âge (dans le cas de certaines usines), et la structure des produits ou de l'industrie; ce peut être aussi des codes qui servent à indiquer, entre autres choses, la création ou la disparition d'une entreprise et les changements de propriété. Ces codes servent au raccordement longitudinal des entreprises et à la définition des liens de propriété entre ces entreprises.

La plupart des données recueillies pour chaque usine concernent les intrants ou les extrants de l'usine. Pour des raisons d'économie d'espace et de temps, nous ne faisons pas ici une description détaillée des variables concernées par ces données. Toutefois, on trouvera une telle description dans le document technique de la LRD, disponible auprès du CES, qui est l'organisme chargé de la mise à jour et de l'exploitation de cette base de données. Néanmoins, la liste de variables reproduite dans le tableau 2 illustre assez bien l'étendue de la LRD. En ce qui concerne les intrants, cette base renferme des données sur les principaux facteurs de production : main-d'oeuvre (pour la production ou autre), capital, matières et services achetés.

Les données sur les extrants comprennent la valeur des expéditions enregistrées pour chaque produit; les années de recensement, ces données portent sur les produits identifiés par 7 chiffres et pour l'EAM, elles portent sur les produits identifiés par 5 chiffres. La LRD renferme aussi des données connexes, comme la valeur ajoutée, les recettes diverses, les stocks, la valeur de revente et les recettes tirées de l'exécution de contrats, pour chaque établissement.

Dans les années de recensement, les données sur les prix sont exprimées le plus souvent sous forme de valeurs unitaires.[8] Pour les années intercensitaires, la LRD ne contient pas les données de quantité normalement nécessaires pour calculer les valeurs unitaires. Par conséquent, lorsqu'on veut exprimer des valeurs en prix constants dans l'estimation de la fonction de production, il faut recourir aux séries de prix par industrie. Ces séries, construites à partir de données du Bureau of Labor Statistics (BLS), sont publiées par le Département du commerce des É.-U. Plusieurs spécialistes s'en servent pour exprimer des valeurs en prix constants.[9]

3. RECHERCHE EFFECTUÉE À L'AIDE DE LA LRD ET DIMENSION TEMPORELLE

Le programme de recherches du CES met en valeur des projets qui exploitent les caractéristiques longitudinales des usines et des entreprises. De nombreux projets sont axés sur la mesure. Ils posent des séries appréciables de critères qui peuvent servir ensuite à des tests d'hypothèses plus approfondis. Mentionnons à titre d'exemple les études de Dunne, Roberts et Samuelson (1988, 1989b), qui portent sur les créations et les disparitions d'entreprises et les flux bruts de la main-d'oeuvre respectivement. Dans les deux cas, on a utilisé des panels de 5 ans formés à l'aide de données de recensement de la LRD. Mentionnons aussi l'étude de Davis et Haltiwanger (1989), qui élaborent de nouvelles mesures de la variation brute et nette de l'emploi sur une base annuelle. D'autres études, réalisées au CES, ont pour objet de tester des hypothèses particulières. Sur ce plan, diverses études, qui analysent l'effet d'un changement de propriété sur le rendement de l'usine ou de l'entreprise, exploitent la structure longitudinale de la LRD. Citons, à titre d'exemple, les articles de McGuckin et Andrews (1988) et de Lichtenberg et Siegel (1988, 1989a, 1989b).

Dans la présente section, nous ne cherchons pas à présenter un tableau complet de la recherche qui se fait à l'aide de la LRD. Par exemple, de nombreuses études sur la mesure de la productivité ne sont traitées que superficiellement. (Plusieurs études ont été publiées au cours des deux dernières années et plusieurs autres projets d'envergure sont en voie de réalisation.) Nous voulons plutôt illustrer le genre de recherches pour une chose: l'industrie à laquelle appartient un établissement, la propriété de l'établissement et l'existence de l'établissement (création et disparition) sont des variables endogènes qui ne peuvent être considérées simplement comme des effets fixes invariants dans le temps dans les modèles économétriques.[2]

On ne peut exagérer l'importance des séries de données de panel pour la recherche économique. Sans elles, de nombreuses questions économiques ne peuvent tout simplement pas être analysées. Parmi ces questions, il s'en trouve toute une série qui porte sur le comportement des agents économiques avant et après la mise en oeuvre de programmes particuliers ou sur d'autres changements dans les conditions ou l'environnement de ces agents. Les données de panel offrent aussi un moyen unique de calculer des micro-mesures de variation brute, ce qui est rarement possible avec des données agrégées.[3]

Les résultats d'études récentes faites par le CES donnent à penser que les mesures de variation brute peuvent être aussi utiles que les mesures de variation nette dans l'analyse de nombreuses questions. Par exemple, des études récentes sur la rotation de la main-d'oeuvre révèlent que les flux bruts de la main-d'oeuvre sont une mesure importante tant au point de vue chronologique (cycles économiques) que transversal (établissements et industries). Des études portant sur la création et la disparition d'entreprises ou d'usines aboutissent aux mêmes conclusions. Dans cet article, nous analysons une mesure de la rotation dans les marchés industriels afin d'évaluer l'importance des mesures de flux brut par rapport aux mesures de flux net normalement utilisées dans l'analyse.

Bien que beaucoup de choses nous échappent encore, nous avons plusieurs raisons de croire que les mesures de variation brute auront des effets économiques appréciables. Premièrement, tout changement implique nécessairement la mise en oeuvre de ressources et les mesures de variation brute permettent justement d'évaluer les coûts afférents et d'en saisir le sens. Deuxièmement, les observations voulant qu'un changement de propriétaire influe sur le rendement de l'entreprise donnent à penser que les mesures de flux brut apportent des renseignements précieux sur la compétitivité.

Une troisième raison d'analyser les variations brutes est qu'elles permettent d'établir si les variations globales sont engendrées par un petit ou un grand segment d'entités économiques. Les personnes responsables de l'élaboration des politiques doivent connaître l'étendue des forces qui sont à l'origine des variations globales. Dans chaque cas, l'utilisation de panels longitudinaux sera nécessaire.

Comme la LRD est relativement nouvelle, il serait utile d'en faire une brève description pour les besoins de l'analyse. Du même coup, nous pourrons en faire l'évaluation comme source de données de panel.

2. LA LRD

On a construit la LRD en raccordant les enregistrements d'établissement tirés du recensement des manufactures (RM), qui a lieu à tous les 5 ans, et de l'enquête annuelle sur les manufactures (EAM). A l'heure actuelle, la LRD compte bien au-delà de 2 millions d'enregistrements qui renferment des données sur plus de 800,000 établissements différents pour la période de 1963 à 1986. Lorsque les résultats du recensement de 1987 seront ajoutés à la base, le nombre d'établissements pourrait dépasser le million.

Le tableau 1 donne le nombre d'établissements compris dans la LRD à chaque année. A chaque année de recensement (1963, 1967, 1972, 1977 et 1982), la base contenait au-delà de 300,000 établissements dont environ les deux tiers avaient participé directement à l'enquête. Les autres, c'est-à-dire ceux dont les données avaient été tirées de dossiers administratifs, étaient de petits établissements (comptant pour la plupart moins de cinq employés) qui influaient peu sur les totaux de l'industrie. Dans les années intercensitaires, la LRD a déjà compté environ 70,000 établissements (période 1973-1978); depuis 1979, année où l'EAM a été remaniée en profondeur, elle en compte environ 55,000.

Selon le plan de l'EAM, la probabilité d'échantillonnage d'une usine est liée directement à sa taille. Cette relation est toutefois complexe. Les gros établissements, ceux qui comptent plus de 250 employés, sont échantillonnés avec une probabilité égale à un. Les établissements plus modestes (ceux qui comptent entre 10 et 250 employés) sont échantillonnés avec une probabilité proportionnelle à la taille de l'effectif sauf que les établissements échantillonnés dans un panel donné ont moins de chances d'être échantillonnés dans le panel suivant[4], cela dans le but de réduire le fardeau de réponse pour les petits établissements. Les panels sont renouvelés à tous les 5 ans afin surtout d'obtenir des estimations précises d'agrégats économiques comme les expéditions.[5]

2.1 Plan transversal

Les données de la LRD proviennent d'enquêtes et de recensements à plan et à traitement transversal. Même si l'étape du contrôle comprend des valeurs d'années antérieures, il y a peu de contrôles fondés sur la chronologie. Pour illustrer le caractère transversal du plan de sondage, précisons que lorsqu'un gros établissement fabrique toute une série de produits différents, le plan considère plusieurs établissements en un. Cela a pour effet d'accroître la précision des agrégats économiques dans l'échantillon mais de rendre moins précis les raccordements d'établissement dans le temps du fait qu'il devient plus difficile de suivre l'évolution de chaque usine.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

DONNÉES ÉCONOMIQUES LONGITUDINALES AU U.S. BUREAU OF THE CENSUS: UN REGARD NOUVEAU SUR DE VIEILLES QUESTIONS

R.H. McGuckin[1]

RÉSUMÉ

Cet article a un double objectif. En premier lieu, il vise à illustrer l'importance des données de panel à l'aide d'exemples puisés dans les études en cours où est utilisée la base de données longitudinales (Longitudinal Research Database -- LRD) du U.S. Bureau of the Census. Bien que la LRD ne soit pas le résultat d'une "vraie" enquête longitudinale, elle permet de constituer des séries de données de panel (équilibré ou non) pour les établissements, les entreprises et les domaines d'activité. En deuxième lieu, cet article vise à intégrer les résultats d'études récentes à la LRD et à déduire des conclusions sur l'importance des micro-données longitudinales pour la recherche économétrique et l'analyse chronologique. Du reste, l'intérêt des données de panel tient au fait qu'il s'agit à la fois de micro-observations et d'observations chronologiques. Par ailleurs, cela nous amène à nous demander pourquoi les données de panel sont indispensables pour comprendre et interpréter l'évolution chronologique des données agrégées produites à la suite de recensements ou d'enquêtes à caractère transversal. Nous allons voir aussi dans cet article que les hypothèses classiques de l'homogénéité peuvent s'avérer invalides dans un grand nombre d'applications. En particulier, l'industrie à laquelle appartient un établissement, la propriété de l'établissement et l'existence de l'établissement (création et disparition) sont des variables endogènes qui ne peuvent être considérées simplement comme des effets fixes invariants dans le temps dans les modèles économétriques.

MOTS CLÉS: Longitudinal; données de panel; LRD; micro-données.

1. INTRODUCTION

"Tu ne peux pas toujours obtenir ce que tu veux mais si tu tentes ta chance, tu obtiens parfois ce qu'il te faut." (TRADUCTION) (Let It Bleed, 1969, Mick Jagger et Keith Richards)

Cet article a un double objectif. Il vise premièrement à illustrer l'importance des données de panel à l'aide d'exemples puisés dans les études en cours où est utilisée la base de données longitudinales (Longitudinal Research Database -- LRD) du U.S. Bureau of the Census. Une série de données de panel est une série formée de multiples observations chronologiques sur des entités économiques. Par exemple, une série de données de panel pour établissement pourrait être constituée d'observations concernant les expéditions des diverses usines d'une période à l'autre. En revanche, les séries chronologiques sont normalement constituées d'observations relatives à un agrégat économique pour des périodes successives, par exemple les expéditions totales d'une industrie ou le revenu national. L'intérêt des données de panel tient au fait qu'il s'agit à la fois de micro-observations et d'observations chronologiques.

Bien que la LRD ne soit pas le résultat d'une "vraie" enquête longitudinale, elle permet de constituer des séries de données de panel (équilibré ou non) pour les établissements, les entreprises et les domaines d'activité.[1] Elle permet aux chercheurs de réaliser de nombreuses études fondamentales jugées irréalisables auparavant. C'est là que la seconde moitié de l'extrait de la chanson de Mick Jagger et Keith Richards prend tout son sens.

En second lieu, cet article vise à intégrer les résultats d'études récentes à la LRD et à déduire des conclusions sur l'importance des micro-données longitudinales pour la recherche économétrique et l'analyse chronologique. L'analyse porte plus spécialement sur le comportement des entreprises et des établissements. Elle nous amène à nous demander pourquoi les données de panel sont indispensables pour comprendre et interpréter l'évolution chronologique des données agrégées produites à la suite de recensements et d'enquêtes à caractère transversal.

La plupart des modèles économiques reposent sur des théories qui concernent le comportement des agents économiques pris individuellement. L'estimation et l'inférence fondées sur des données agrégées s'accompagnent d'hypothèses sur l'homogénéité des entités qui forment l'agrégat. On pourrait supposer, par exemple, que la répartition des entités en fonction d'une variable particulière comme l'efficience ou la classification industrielle ne varie pas dans le temps. Nous allons voir dans cet article que des hypothèses de ce genre peuvent s'avérer invalides dans de nombreuses applications. Les observations nous amènent à constater

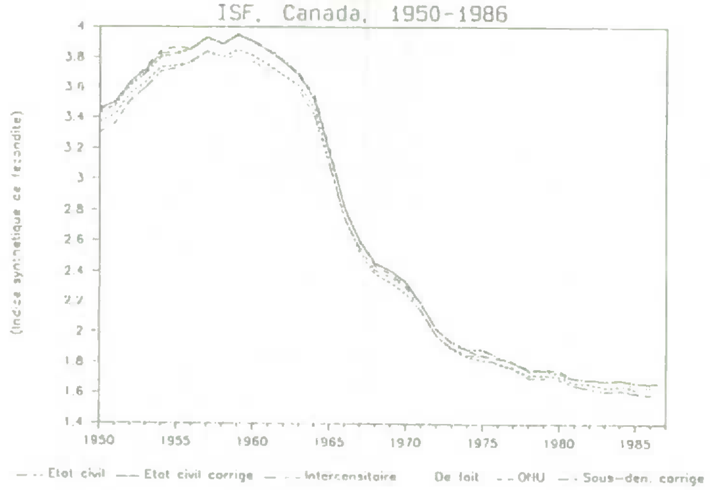
[1] R.H. McGuckin, chef, Center for Economic Studies (U.S.), Bureau of the Census, Washington, D.C. 20233. Cet article ne reflète pas nécessairement la position du U.S. Bureau of the Census. L'auteur remercie Timothy Dunne et John Haltiwanger pour lui avoir communiqué leurs commentaires et avoir accepté de participer à de nombreux échanges fructueux. Il remercie également Stephen Andrews, Scott Schuh, James Monahan, Sang Nguyen et Mary Streitwieser pour leurs précieux commentaires. Enfin, il tient à exprimer sa reconnaissance à Stephen Andrews et à Robert Bechtold, qui ont compilé les données, de même qu'à Janice Bryant et à Terry Norris, qui ont accompli un magnifique travail à la machine à écrire.

SECTION 7

ÉCONOMÉTRIE

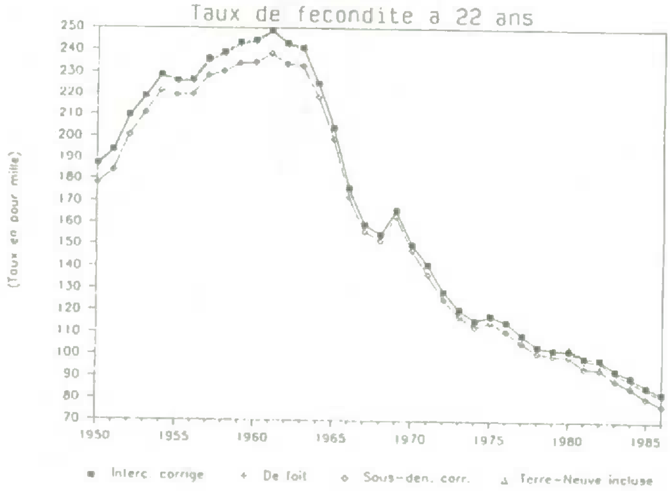
Graphique I

ISF, Canada, 1950-1986



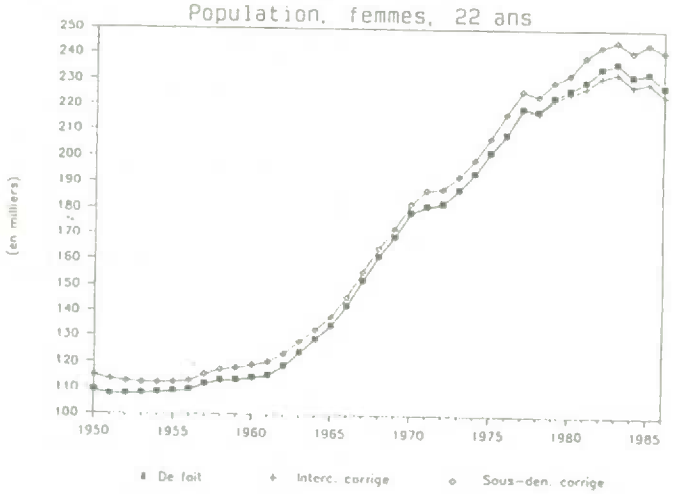
Graphique II

Taux de fecondite a 22 ans



Graphique III

Population, femmes, 22 ans



possibles, et ce dans différentes situations. Bien sûr, une telle opération est impossible et se doit donc d'être menée directement par chaque chercheur en début de travail.

5. CONCLUSION

Dans toute série chronologique, les discontinuités potentielles sont nombreuses. Elles vont de l'arrêt de la parution d'une variable à la modification de la qualité des données. Les réactions face à celles-ci varient également d'un chercheur à l'autre: abandon, restriction de l'étude ou correction des informations lacunaires.

En fait, il n'y a pas de recette miracle. Le chercheur est seul juge en matière de discontinuités. Chaque travail est particulier en ce sens que non seulement le thème étudié importe dans la décision finale, mais également l'environnement: les sources disponibles, la main-d'oeuvre et même les ressources financières, les contacts et le poste du chercheur. Celui qui a accès à l'information de base (celle des formulaires ou questionnaires) n'est pas dans la même situation que celui qui travaille seul et qui ne peut obtenir des tableaux détaillés qu'à prix fort. Et ce, même s'ils travaillent sur le même sujet.

L'étude de l'évolution de l'indice synthétique de fécondité menée à titre d'exemple, nous a permis de dégager certains avenues de décision qui bien qu'incomplètes montrent bien que l'impact des discontinuités n'est pas toujours énorme. Il varie selon le niveau de l'erreur, le type de la discontinuité, l'élément étudié et son niveau d'agrégation, le poids de la population touchée par la discontinuité dans la population totale ciblée. Points importants à ne pas oublier: le niveau géographique et surtout, la taille de la population étudiée.

Quoiqu'il en soit, il importe de ne plus regarder les séries chronologiques de données du même oeil et d'ajouter la prudence, l'attention aux méthodes utilisées. Enfin, n'attribuons pas trop d'importance à des différences minimes dans le temps mais mettons plutôt l'emphase sur les variations majeures et les tendances soutenues.

REMERCIEMENTS

L'auteure aimerait remercier pour leurs précieux conseils et suggestions, M. Jean Dumas et Mesdames Margaret Michalowski et Jeannine Perreault. L'assistance technique de Mesdames Monique Charbonneau, Denise Laflamme et Lili Stevenson Roy a également été très appréciée.

BIBLIOGRAPHIE

Bergstrom, T., et Lam, D. (1989), "Recovering event histories by cubic spline interpolation," Mathematical population studies, 1, 327-355.

Henry, L., et Blayo, Y. (1975), "La population de la France de 1740 à 1860", Population, 30, 71-122.

Islami, H. (1983), "La population albanaise de Yougoslavie: accroissement numérique et répartition spatiale", Population, 38, 166-173.

Kalbach, W.E., et McVey, W.W. (1971), The demographic bases of Canadian society, Toronto: McGraw-Hill Ryerson Limitée.

Lowe, R.J. (1987), "Comparing 1981 and 1986 Census Labor Force employment and unemployment data," New Zealand population review, 13, 27-34.

Meslé, F., et Vallin, J. (1981), "La population des établissements psychiatriques: évolution de la morbidité ou changement de stratégie médicale", Population, 36, 1035-1068.

Monnier, A. (1982), "La conjoncture démographique: l'Europe et les pays développés d'outre-mer", Population, 37, 911-940.

———— (1986), "La conjoncture démographique: l'Europe et les pays développés d'outre-mer", Population, 41, 823-845.

Munoz-Perez, F., et Tribalat M. (1984), "Mariages d'étrangers et mariages mixtes en France. Évolution depuis la première guerre", Population, 39, 427-462.

O.N.U.(1979), Annuaire démographique 1978, supplément rétrospectif. New York: Nations-Unies.

———— (1987), Annuaire démographique 1986, sujet spécial: statistiques de la natalité, New York: Nations-Unies.

Perrégaux, J.-C. (1983), "La hausse de la mortalité infantile en U.R.S.S.: mythe ou réalité", Population, 38, 1050-1055.

Rallu, J.-L. (1986) "Descendance des générations françaises et probabilités d'agrandissement", Population, 41, 763-802.

Roussel, L. (1983), "Les ménages d'une personne: l'évolution récente", Population, 38, 995-1015.

Statistique Canada (1987), Méthodes d'estimation de la population, publication no. 91-528F au catalogue, Ottawa: Ministère des approvisionnements et services.

———— (annuel), Statistiques de l'état civil, Ottawa.

Taieb, J. (1982), "Évolution et comportement démographique des Juifs de Tunisie sous le protectorat français (1881-1956)", Population, 37, 952-958.

Vallin, J. (1983), "Tendances récentes de la mortalité française", Population, 38, 77-105.

important que pour l'I.S.F., l'écart moyen de 1977 à 1986 étant de 4.35%. Quant à l'effet de l'ajout de la province de Terre-Neuve, il est faible mais néanmoins visible, les taux de cette série étant les plus élevés en 1980 et 1981, seules années de disponibilité.

Le taux de fécondité par âge, comme l'I.S.F., est un rapport liant deux sources. Nous allons maintenant mesurer l'impact de certaines discontinuités sur la population féminine à 22 ans. Les séries précédentes à l'exception de celle incluant Terre-Neuve seront l'objet de la comparaison. Le graphique III illustre les tendances. Encore une fois, elles sont similaires mais la position des courbes est à l'opposé de celle montrée précédemment. Rien de surprenant, la population servant de dénominateur dans l'I.S.F. et le taux par âge, d'où l'inversion. L'évolution des estimations intercensitaires de population se juxtapose à celle de la population de fait-de jure, du moins jusqu'en 1976, date à laquelle la seconde s'accroît plus rapidement. En 1986, les deux séries sont distantes de 1.6%, écart tout de même respectable. L'effet de la correction du sous-dénombrement est intéressant. Non seulement, est-il apparent ici que l'écart se creuse entre les deux autres séries et la série corrigée du sous-dénombrement mais, alors que les autres semblent montrer une baisse de la population féminine âgée de 22 ans depuis 1983, la série corrigée laisse présager plutôt d'une certaine stagnation de l'effectif. En 1986, l'écart entre estimations intercensitaires et population corrigée du sous-dénombrement est de 5.8%.

TABLEAU 1

Impact de certains types de discontinuités sur trois indicateurs démographiques

Discontinuité	Écart mesuré (%)		
	I.S.F.	Taux de fécondité à 22 ans	Population féminine à 22 ans
. méthode de calcul de l'indice	2.5	n.d	n.d.
. méthode de calcul des estimations intercensitaires	négligeable	n.d	n.d.
. type de population: postcensitaire ou intercensitaire	négligeable	n.d	n.d.
. population cible: de jure ou de jure-de fait	1.2	1.0	1.6
. territoire	négligeable	1.7	n.d.
. sous-dénombrement aux recensements	4.0	4.4	5.8

n.d. = non disponible

Pour résumer quelque peu les commentaires précédents, il nous paraît utile de réunir les observations relatives à l'impact des différentes discontinuités étudiées dans un seul tableau, le tableau 1. Même s'il ne concerne que certaines mesures démographiques bien particulières et une sélection de discontinuités, ce tableau est riche en enseignements. Il nous montre d'abord que plus le niveau d'agrégation est élevé et éloigné de la source de discontinuité, plus l'impact de la discontinuité est faible. Par exemple, l'impact d'une même discontinuité est beaucoup plus forte sur l'élément touché directement, la population en l'occurrence, que sur tout indice qui utilise cette population, et là encore, l'impact diminue si l'indice calculé n'utilise pas que la population de la catégorie à fort degré d'erreur. Le niveau de l'impact dépend donc du degré d'agrégation de l'élément étudié par rapport au niveau d'arrivée de la discontinuité. Cependant, le poids de ce dernier niveau dans l'élément mesuré importe aussi. Nous pouvons affirmer sans grand risque d'erreur que l'impact du sous-dénombrement est certainement plus grand sur l'I.S.F. que sur l'espérance de vie à la naissance, les naissances survenant à un âge où le sous-dénombrement est élevé, au contraire des décès. Le tableau 1 nous rappelle également que certaines discontinuités n'ont pas ou peu d'effets, l'impact de chacune étant intrinsèquement lié aux caractéristiques-mêmes de la discontinuité. Ici, la modification du territoire joue peu parce que la population de Terre-Neuve est faible dans l'ensemble canadien. Qu'en aurait-il été si l'Ontario avait été la province omise?

L'exemple présenté ici concerne des indices qui ne peuvent que croître ou décroître sans être négatifs: seul le format peut varier non le sens. Le phénomène migratoire est différent, et l'impact des discontinuités peut aller jusqu'à modifier le sens du mouvement net total. L'étude de tels phénomènes demande donc d'autant plus de précautions.

Il aurait été désirable de dresser une liste exhaustive des discontinuités identifiées au premier chapitre et d'en citer, avec chiffres à l'appui, l'impact sur tous les indices démographiques

Il convient de souligner que ces six discontinuités ne sont pas les seules dont souffrent les données: l'enregistrement à l'état civil est provincial tandis que le recensement (et les estimations qui en découlent) est du ressort du fédéral; la définition de "naissances vivantes" a été modifiée en 1959 et l'application de la nouvelle définition s'est faite à date variable selon les provinces; enfin, depuis 1981, le recensement du Canada ne dépend plus du travail des énumérateurs puisque chaque ménage complète lui-même son questionnaire et le retourne ensuite par la poste. Nous nous limiterons cependant à l'étude de l'impact des six nommées précédemment, d'une part parce que celles négligées touchent l'ensemble de nos séries de la même façon, et d'autre part, parce que leur impact est beaucoup plus complexe à mesurer.

Le graphique I montre que quelle que soit la série identifiée, la tendance de l'évolution de l'indice synthétique de fécondité de 1950 à 1986 est la même: hausse constante jusqu'en 1957, puis chute rapide jusqu'en 1973 suivie d'une baisse beaucoup moins accentuée. Des différences apparaissent toutefois pour les séries des Nations-Unies et corrigée du sous-dénombrement. L'examen détaillé des tendances d'année en année montre parfois quelques variations. De 1983 à 1984, toutes les séries montrent une légère hausse à l'exception de celle basée sur la population "de jure-de fait" qui indique plutôt une stagnation. Le même phénomène existe en 1974-75.

L'effet des six discontinuités identifiées précédemment peut être mis en lumière par la comparaison deux à deux des séries dans lesquelles un seul élément diffère. L'impact de la méthode de calcul de l'I.S.F. semble être important: le passage de taux par âges simples aux taux par groupes d'âges quinquennaux diminue de beaucoup le niveau de l'indice (2.5% environ), et ce durant toute la période. La comparaison des deux séries basées sur les estimations intercensitaires (révisées ou non) ne montre, à peu de choses près, aucun écart, sinon de très légers. Le changement de méthode de calcul n'a donc que peu d'effet. Ce n'est guère surprenant puisque, par définition, et ce, quelle que soit la mèthode utilisée, les estimations intercensitaires découlent de deux recensements encadrants, qui eux, ne varient pas. L'impact de l'utilisation des séries postcensitaire ou intercensitaire est également négligeable. Là encore, ce n'est pas une surprise puisqu'il a été démontré que la différence, au niveau canadien, entre les deux types de population, est généralement inférieure à 1% (Statistique Canada 1987). Le changement dans les populations cibles, de "de jure" à un "de jure-de droit" est un peu plus sensible et ce, surtout sur les dernières années. Cela tient à nos hypothèses de correction (basées sur nos connaissances fragmentaires): les résidents temporaires de plus d'un an au Canada (réfugiés en attente de statut, étudiants, détenteurs de permis de travail, diplomates, visiteurs) ne cessent d'augmenter. La correction simulée, enfin, du sous-dénombrement des différents recensements de population, est, elle, d'effet majeur, et ce, particulièrement entre 1950 et 1965, et depuis 1976. Cette correction repose sur les seules données connues relatives au sous-dénombrement, publiées dans les documents relatifs à chacun d'entre eux. Les méthodes d'obtention ont varié avant 1961, mais, depuis cette date, celle utilisée est la contre-vérification des dossiers, opération qui recherche parmi les personnes recensées un échantillon de personnes qui auraient dû l'être. La taille de l'échantillon n'a cessé d'augmenter depuis cette date mais les niveaux d'erreur demeurent élevés. Nous rappelons donc qu'il ne s'agit ici que d'une simulation. L'impact entre 1950 et 1965 est important puisqu'il modifie l'indice d'en moyenne 4.4%. Le niveau élevé du sous-dénombrement entre 20 et 24 ans se répercute donc ici tout comme celui du recensement de 1961 (3.3%). L'impact entre 1976 et aujourd'hui est un peu plus faible (3.5%), même s'il va en s'accentuant. Cette différence entre les deux périodes ne tient pas à une diminution du taux de sous-dénombrement dans le temps, mais plutôt à une augmentation de l'âge moyen des mères à la naissance, les femmes de 25-29 ans étant moins sous-dénombrées que celles de 20-24 ans. L'indice révisé pour inclure Terre-Neuve est absent du graphique puisqu'il n'a été calculé que pour deux années. La correction n'a cependant aucun impact, les indices étant identiques à ceux basés sur les estimations intercensitaires révisées. Il est vrai que la population de Terre-Neuve est faible eu égard à la population canadienne totale.

À quelques exceptions près, donc, l'impact des discontinuités sur un indice synthétique du type de celui étudié précédemment est plutôt faible. La prise en compte, dans un tel indice, d'un grand nombre de taux différents, dilue en quelque sorte l'erreur qui pourrait entacher plus particulièrement l'un d'entre eux. Quel serait l'impact de certaines de ces discontinuités sur des taux spécifiques, à l'âge par exemple où elles sont les plus fortes? L'impact du sous-dénombrement étant important sur l'I.S.F., et sachant qu'il touche particulièrement les 20-24 ans, nous allons étudier l'évolution du taux de fécondité générale à 22 ans, de 1950 à 1987. Nous ne pourrons commenter que quatre séries, les trois séries officielles publiées ne fournissant aucune information sur les taux par âges simples.

La graphique II illustre l'évolution comparative des séries basées sur les intercensitaires révisées, la population de fait-de jure, le sous-dénombrement corrigé et l'ajout de la province de Terre-Neuve. Comme pour l'I.S.F., les deux premières sont quasi-identiques, même si la seconde est légèrement inférieure en fin de période. En fait, l'écart entre elles est moindre que pour l'I.S.F. et cela tient, certainement en partie, au fait que la population féminine ajoutée dans la population de fait-de jure a 27.7 ans d'âge médian et que l'impact majeur sur les taux n'est pas à 22 ans. L'impact de la correction du sous-dénombrement est cependant, tel que prévu, plus extérieure. La qualité globale de l'information publiée, par exemple, ne lui est connue que grâce aux études de l'organisme responsable. Bien sûr, il peut mener lui-même quelques travaux comparatifs, mais pour certaines variables (le sous-dénombrement par exemple), le chercheur est tributaire de décisions d'autrui. Et alors, les résultats fournis peuvent ne pas répondre à ses attentes et n'être, comme c'est souvent le cas, qu'inutilisables car trop superficiels. Ce problème n'est cependant pas insoluble. Quelques-uns des exemples présentés montrent bien que, parfois, il n'est pas nécessaire de corriger les données, l'impact des discontinuités sur les résultats étant négligeables. Reste donc à mesurer cas par cas cet impact.

4. IMPACT DES DISCONTINUITÉS

L'impact des discontinuités peut être mesuré de façon fort simple par l'utilisation de tests de sensibilité. Pour ce faire, il s'agit de poser sur nos données une série d'hypothèses quant à leur niveau d'erreur, de les corriger pour en tenir compte et de voir l'impact de ces corrections sur les résultats. Lorsque celui-ci est négligeable, tout va pour le mieux. Dans le cas contraire, la solution serait de corriger les données si le niveau réel de l'erreur est connu. Sinon, la décision finale de poursuivre l'étude appartient au chercheur: il est seul juge. Cette décision se prendra cependant en toute connaissance de cause et s'il décide de poursuivre, il se devra de mettre en garde le lecteur contre les conclusions hâtives dues à de faibles variations dans la série et, pourquoi pas, d'en expliquer l'origine possible.

Ne soyons pas pessimiste cependant. Nous allons clore cet article sur un essai d'évaluation de l'impact de certaines formes de discontinuités sur différents indices démographiques. Pour ce faire, nous allons utiliser l'exemple concret de l'évolution de l'indice synthétique de fécondité (I.S.F.) au Canada de 1950 à 1986. Cet indice s'obtient, par année, de la somme des taux de fécondité générale par âge des femmes. Les taux de fécondité générale font le rapport des naissances de femmes d'âge x sur l'effectif total de ces femmes.

Plusieurs choix s'offrent à nous quant aux données à utiliser. Les indices synthétiques de fécondité sont calculés régulièrement et apparaissent dans diverses publications officielles. Au niveau international, l'annuaire démographique des Nations-Unies fournit les taux de fécondité générale par groupes quinquennaux d'âge, au Canada, pour la période sélectionnée, lesquels multipliés par 5 et sommés fournissent l'indice cherché. Ces taux excluent les naissances de la province de Terre-Neuve mais sont basés sur la population totale canadienne. Les résidents canadiens temporairement aux États-Unis sont également comptés (O.N.U. 1979; 1987). Au niveau national, Statistique Canada (et antérieurement le Bureau fédéral de la statistique), publie annuellement les données de l'état civil et quelques indices simples qui en sont tirés, dont l'I.S.F. Jusqu'en 1972, deux séries sont disponibles: les indices tels que publiés annuellement et basés sur les estimations postcensitaires de population; les indices révisés après chaque recensement et donc basés sur les estimations intercensitaires de population. Depuis 1972, cette dernière série n'existe plus. Nous pourrions également décider de calculer nous-mêmes ces indices au moyen des données définitives de naissances et des estimations intercensitaires de population, révisées. Terre-Neuve ne disposant pas de l'information sur les naissances par âge de la mère, elle est exclue du calcul des indices canadiens précédents. Nous pourrions estimer ces naissances et recalculer l'I.S.F. en incluant cette province. Certains tests de sensibilité (simulations) seraient également possibles. Les populations cibles de l'enregistrement à l'état civil et au recensement étant différentes (état civil = population habituellement résidente + certains Canadiens à l'étranger, dont les militaires et diplomates; recensement = population "de jure"), nous pourrions choisir d'estimer l'écart entre elles, en majorer l'effectif de femmes soumises au risque de procréer et recalculer les taux. Enfin, puisqu'il est connu que les recensements canadiens sont sous-dénombrés, nous pourrions également, sur la base des taux de couverture connus par la contre-vérification des dossiers (bien que peu fiables, la marge d'erreur étant très importante), en corriger la population de fait augmentée précédente.

Nous aurions donc, pour une même analyse, sept séries différentes de données. Leur étude comparative nous permettrait de mesurer ou du moins d'évaluer l'impact de leurs différences (lesquelles sont en fait autant de discontinuités dans les séries) sur l'indice étudié et son évolution dans le temps. Ces discontinuités sont les suivantes:

1. Changement dans la méthode de calcul de l'indice: les Nations-Unies utilisent des taux par groupes d'âges quinquennaux; les autres séries, des taux par années simples d'âge.
2. Changement dans la méthode de calcul des populations intercensitaires: estimations révisées après chaque recensement pour celles utilisées par Statistique Canada; estimations révisées après coup, par l'utilisation de la même méthodologie sur toute la période.
3. Changement dans le type de population estimée: population postcensitaire ou intercensitaire selon les deux séries publiées par Statistique Canada.
4. Changement dans la population cible au dénominateur: population de jure règle générale, mais population ciblée par l'état civil pour les deux dernières séries présentées.
5. Changement dans le territoire: inclusion ou non de Terre-Neuve.
6. Changement dans la qualité des données de population: ajustement ou non pour le sous-dénombrement.

Il n'est pas inhabituel, dans la littérature, de chercher en vain des commentaires sur la qualité des données utilisées et les ruptures possibles sur le long terme. Bien sûr, certaines études reposent sur des informations maintes fois utilisées et bien connues. Ce n'est malheureusement pas toujours le cas. Comment savoir alors s'il y a des discontinuités ou s'il y en avait et, dans un tel cas, comment elles ont été corrigées. Le lecteur peut s'aider, pour asseoir son opinion, de l'énumération des sources utilisées. Cette solution n'est pas nécessairement mauvaise: tout dépend du nombre, de la gravité et du sens des discontinuités.

Il importe donc qu'avant de commencer son analyse, le chercheur détermine la qualité globale de sa série et sa capacité à mesurer l'évolution de l'élément choisi. Certains le font et, s'ils décident par la suite de conserver leurs données telles qu'elles, c'est en pleine connaissance de cause. Par exemple, Roussel (1983), dans une analyse de l'évolution des ménages aux Pays-Bas depuis 1945, note un changement de concept. Il en mesure l'impact sur les ménages touchés et réalise qu'il est peu important (à peine 2.5%). Il l'est d'autant moins lorsque mis en relation avec les mouvements de grande amplitude qu'ont connu les ménages sur la période. L'auteur conclut donc que "ces biais sans être négligeables, n'invalident pas les conclusions" (Roussel 1983, page 997). La décision de ne pas ajuster les données est donc basée sur une mesure d'impact des discontinuités. D'autres la prennent sur la base du temps dont ils disposent ou des coûts qu'entraîneraient les corrections éventuelles. Néanmoins, ils mettent en garde le lecteur contre les conclusions hâtives découlant de faibles variations et énumèrent tous les cas de discontinuités. L'exemple type est le livre de Kalbach et McVey, paru en 1971.

Parmi ceux qui modifient leur plan de travail original suite à des discontinuités, il y a des restrictions dans la durée de temps retenue ou un moindre détail dans les classifications; des abandons; des corrections. Nombreux sont les chercheurs qui faute de temps, d'argent ou de techniques appropriées, décident de limiter quelque peu le détail de leur analyse, en terme de durée, de caractéristiques ou de classification. J. Vallin (1983) n'a, par exemple, fait porter sa description de l'évolution de la mortalité française que sur la période 1950-1978, évitant du même coup les distorsions probables engendrées par les 5^{e} et 9^{e} révisions du code international des maladies. Il était néanmoins confronté à celles des 6^{e}, 7^{e} et 8^{e} révisions qu'il a tenté de limiter au maximum par l'adoption de niveaux de classification plus globaux, similaires dans les trois séries. Certains chercheurs n'ont d'autre choix que d'abandonner l'étude envisagée. Citons, à titre d'exemple, R.J. Lowe (1987) qui a dû renoncer à étudier l'évolution de l'emploi à plein temps en Nouvelle-Zélande à partir des recensements de 1981 et 1986, devant la modification de la ligne de partage entre emplois à temps plein et partiel entre les deux dates.

Tous, cependant, ne doivent pas abandonner ou modifier leur idée première. Bon nombre cherchent à faire le lien entre les données dans le temps, à transformer soit la série originale, soit la série d'appoint ou, à tout le moins à améliorer la qualité de l'information mesurée. Il n'y a pas de méthodes statistiques éprouvées pour ce faire. Il n'y a que des exemples qui éventuellement pourront servir de suggestions aux chercheurs confrontés à des problèmes identiques. Ces exemples nombreux concernent la majeure partie des discontinuités identifiées. Nous avons sélectionné ceux qui nous ont paru les plus pertinents et novateurs. Les méthodes les plus complexes ne sont pas nécessairement les meilleures. Parfois la réalité correspond plus ou moins à une forme simple, facile à approcher. Par exemple, F. Munoz-Perez et M. Tribalat (1984), dans leur étude de l'évolution du nombre de mariages mixtes en France de 1910 à 1982, ont, pour combler les lacunes de 1932 à 1942, simulé les événements par une simple interpolation linéaire, basée sur des données fragmentaires. D'autres corrigent les effectifs connus au prorata de la population manquante (Meslé et Vallin 1981), changent leur niveau d'analyse, préférant à une étude globale qui aurait camouflé les incohérences, une étude régionale qui les met en lumière mais uniquement là où elles sont vraiment (Perrégaux 1983), ou encore, remplacent les nombreux actes manquants au moyen d'une procédure simple de sélection aléatoire parmi ceux disponibles (Henry et Blayo 1975). Une autre solution est de faire appel à des sources tierces qui, bien que négligées dans l'analyse principale, servent à corriger les écarts ou assurer les passages dans le temps. J.-L. Rallu (1986), par exemple, confronté au changement du concept de "rang de naissance" sur la période qu'il voulait étudier, s'est servi des résultats de deux enquêtes - familles réalisées dans le même intervalle de temps. Ces dernières fournissaient l'information selon l'ancienne et la nouvelle définitions et donc, M. Rallu a pu calculer des coefficients par périodes et groupes quinquennaux d'âges, lesquels ont été appliqués à la série principale de données. D'autres auteurs font plutôt intervenir une méthodologie mathématique élaborée. Pour preuve, ne citons que l'exemple de Bergstrom et Lam (1989), qui préconisent l'emploi d'interpolations cubiques pour remédier au problème de l'incohérence lors de la translation de données par âges en données par année de naissance.

De ces différents exemples, se dégage une approche précise à la correction éventuelle de séries chronologiques lacunaires. D'abord, et c'est là une étape extrêmement importante, toute décision doit être précédée d'une évaluation précise de la situation: identifier les discontinuités de la série dont on dispose, leur niveau et leur impact sur les résultats. Si l'identification des discontinuités n'est pas trop difficile, il n'en va pas de même de la mesure de leur niveau. Règle générale, le chercheur n'a aucun contrôle sur la base de données puisqu'elle lui est décès et émigrants à la population de base, généralement celle du dernier recensement disponible. La discontinuité totale sur la période tient alors à l'ensemble des discontinuités de chaque source de données, plus à celles les interreliant, pour chacun des points énumérés en A. Ceci vaut aussi, bien sûr, dans le calcul des ratios.

Les discontinuités dans les interrelations entre sources peuvent provenir du changement de la méthode les liant entre elles, mais aussi d'une part de toute modification affectant l'une des sources sans le faire des autres et d'autre part, de tout changement altérant l'écart entre les sources et ce sans qu'en apparence, rien n'ait changé.

Le souci du détail et de la précision pousse les chercheurs à raffiner de plus en plus leurs méthodes de calcul, modifiant du même coup la comparabilité entre les années. À titre d'exemple, soulignons qu'en Roumanie, avant 1975, l'espérance de vie de la population était calculée sur une seule année, alors qu'après cette date, elle l'était sur trois ans (Monnier, 1982). Il convient de rappeler cependant que lorsqu'une méthode, un concept ou une orientation, est modifié par l'organisme statistique diffuseur de données, il est d'usage que soient faites soit une correction rétrospective, soit une présentation durant quelques années de la double série.Ce comportement permet de limiter l'impact des incohérences dans les séries chronologiques. Quant à une modification qui n'affecterait qu'une seule des sources utilisées de pair, les exemples sont nombreux. En voici un: le calcul de taux de mortalité, au Canada, basés d'une part sur les décès (État civil, Statistique Canada), inchangés, et d'autre part sur la population estimée au moyen d'une méthode modifiée (Division de la démographie, Statistique Canada). Enfin, certains changements peuvent ne toucher aucune des sources impliquées mais l'écart entre elles. Cette situation est beaucoup plus évidente par l'examen de cas concrets. Depuis toujours, le calcul des indices de fécondité du Canada nécessite l'utilisation de populations cibles différentes: au numérateur, les naissances proviennent de la population de fait au pays et des Canadiens à l'étranger; au dénominateur, la population est de type "de jure". Ces concepts n'ont changé sur aucun point; par contre, la proportion de la population de fait qui n'est pas de droit ne cesse d'augmenter: réfugiés, étudiants avec visas, détenteurs de permis de travail temporaires et, pourquoi pas, immigrants illégaux. L'accroissement de l'écart entre sources peut donc aller jusqu'à fausser les conclusions d'une analyse en infléchissant les tendances.

En dernier lieu, nous voudrions souligner que ces sources utilisées conjointement deviennent en elles-mêmes une autre source et qu'en ce sens, les mises en garde énoncées en A s'appliquent de la même façon que lors de l'emploi d'une source unique. Par exemple, un indice peut être affecté par la modification de sa population cible, comme à Malte, où, avant 1975, l'indice synthétique de fécondité concernait l'ensemble des femmes résidentes et par la suite, exclut les étrangères.

Nous nous devons également d'insister sur le niveau de complétude des données et leur type. Les estimations de population, par exemple, se distinguent sur trois points. Elles peuvent être soit postcensitaires, i.e. produites à partir du dernier recensement disponible et des modifications des composantes depuis, soit intercensitaires, i.e. basées sur les estimations postcensitaires et les deux recensements encadrants. Leur périodicité peut également différer: elles sont soit trimestrielles, soit annuelles. Enfin, leur degré de finition varie aussi selon qu'elles soient provisoires (3 à 4 mois après la date de référence), mises à jour (8 mois) ou définitives (15 à 20 mois). Le chercheur doit donc s'assurer qu'il a en main les estimations intercensitaires définitives à toute date, ou du moins, autant que faire se peut. Il doit donc, du même coup, rechercher les séries d'indices statistiques nationaux corrigés, basés non plus sur les estimations postcensitaires comme le sont les calculs annuels, mais sur les estimations intercensitaires ou le recensement selon qu'on soit ou non une année de recensement.

2.4 Changements de sources utilisées conjointement, chronologiquement

Une série de données basée sur deux sources ou plus peut souffrir, sur le long terme, outre les coupures ci-haut mentionnées, d'incohérences dues au changement de l'une ou l'autre des sources sinon plusieurs. Les discontinuités potentielles sont alors celles citées en B et en C.

Les discontinuités identifiées ne sont pas toutes de même format et leur impact ne va pas nécessairement dans le même sens. Certaines peuvent simuler une hausse, d'autres une baisse. Le cumul des coupures dans une série chronologique peut donc n'être que léger et ce même si elles sont nombreuses. À l'inverse, une seule d'entre elles peut entraîner des modifications majeures dans les nombres et surtout dans les tendances. Comment le chercheur peut-il s'y retrouver? De quels outils dispose-t-il pour corriger ces discontinuités? Et le fait-il généralement?

3. CORRECTION DES DISCONTINUITÉS

La réaction des démographes devant une série chronologique à faire ou déjà formée n'est pas uniforme. Les comportements vont d'un extrême à l'autre, de ceux qui ne tiennent pas compte des discontinuités à ceux qui le font. Et encore, parmi ces deux groupes, les stratégies varient.

changement de la méthode de correction dans le temps. Les procédures d'ajustement sont extérieures aux réponses elles-mêmes et existent en réponse au souci de confidentialité des répondants et de leurs gouvernants. Statistique Canada, par exemple, ne fournit plus depuis 1971 de tableaux croisés détaillés contenant des cellules quasi-vides. Toutes les données finales sont ajustées par un procédé nommé "arrondissement aléatoire" qui camoufle derrière un "0" ou un "5" des nombres plus petits. La procédure de pondération affecte principalement les enquêtes mais peut aussi être utilisée lors de recensements dont une partie n'est administrée qu'à un échantillon de la population totale. Cette procédure qui extrapole à l'ensemble de la population les résultats de l'échantillon génère ce qu'il est convenu de nommer "erreur d'échantillonnage" ou "erreur aléatoire". Le biais provient de la taille et du mode de tirage de l'échantillon, de la méthode de redressement utilisée et de ses composantes (source de données, méthode d'obtention...). Tout changement dans l'un ou l'autre de ces éléments dans le temps peut affecter la comparabilité des données. Avant de clore ce paragraphe sur les erreurs de traitement, nous voudrions rappeler le rôle de plus en plus important joué par la technique et la machinerie (outils méchanographiques, ordinateurs principaux, mini ou micro-ordinateurs) à cette étape. Leur utilisation systématique à tous les niveaux du processus de traitement a grandement amélioré d'année en année non seulement la gestion des opérations mais également la qualité du traitement, par l'élimination des transferts et des erreurs humaines. Enfin, ajoutons que l'erreur aléatoire ne provient pas que de la procédure de pondération mais peut aussi découler de l'utilisation de petits nombres. Il est connu que plus la taille de la population ciblée est petite, plus l'erreur aléatoire est grande. Dans une analyse des séries chronologiques où la population croît dans le temps, il est bon de se rappeler que la qualité des données croît de pair et ce, d'autant plus, que la population de départ est faible et celle d'arrivée élevée.

La dernière cause d'erreur que nous allons relever ici est celle liée à l'unité de la collecte. Il arrive que la responsabilité de la cueillette d'information sur un territoire soit divisée entre plus petites unités géographiques, les résultats partiels générant le résultat total. Il appert alors que les politiques et procédures d'enregistrement sinon de traitement peuvent varier d'une unité à l'autre, engendrant des incohérences au niveau global.

La majeure partie de ces erreurs affectant la qualité des données ne peut être ou n'est pas mesurée. On en connaît néanmoins l'ordre de grandeur et le degré de variabilité, fonction non seulement des caractéristiques-mêmes de la source ou des répondants mais aussi et surtout, en fait, du temps. Dans une étude de séries chronologiques, ce dernier point s'avère un élément qu'il ne faut pas négliger. En effet, il semble exister une relation directe entre la période de collecte de l'information et sa qualité: la qualité croissant avec la date de la collecte. Sans toutefois aller jusqu'à affirmer que les données anciennes sont inutilisables, il est sûr que l'effort fourni, l'infrastructure mise en place, l'amélioration des techniques, l'augmentation des connaissances, des moyens financiers et de l'expérience surtout, n'ont pu qu'accroître la qualité des données. Or, à ces améliorations s'en sont greffées qui ont servi à mettre en place des techniques plus raffinées de mesure de la qualité de l'information et d'identification de la nature des erreurs. Tout cela pour en arriver à une situation paradoxale où parfois les erreurs semblent augmenter à mesure que les techniques employées s'améliorent!

2.2 Changements de sources dans le temps chronologiquement

L'utilisation d'une seule et même source de données sur une longue période de temps nous impose des contraintes extérieures nombreuses, comme nous venons de le montrer, qui toutes mènent à des discontinuités plus ou moins importantes. La décision de changer la source utilisée ne peut qu'amplifier les problèmes, les discontinuités de l'une se joignant à celles de l'autre. Il est vrai que l'addition des incohérences n'est pas aussi simple que cela: il peut arriver qu'une même modification affecte les deux sources à un même moment et qu'alors, sur ce point du moins, la translation ait peu d'effet. Il se peut également qu'une modification apportée à la source-maître (la première utilisée), peu importe sous quel aspect, soit si importante qu'il vaut mieux faire appel à une autre source, quitte à générer des discontinuités nombreuses mais à tout le moins mineures. Enfin, il arrive qu'on n'ait pas vraiment le choix, l'élément étudié n'étant plus disponible via la première source et ne l'ayant pas été auparavant par la seconde. Quoiqu'il en soit, il importe, tant en ce qui concerne les discontinuités de premier que de deuxième niveau, de faire le point sur les variations totales engendrées par ces coupures sur la période étudiée.

2.3 Changements dans les relations entre sources utilisées conjointement, dans le temps

Il arrive aussi, fréquemment, qu'une analyse quelconque doive utiliser simultanément deux sources de données ou plus. C'est le cas, en particulier, de l'étude de l'évolution de taux, quotients et autres indicateurs sociaux et économiques. Dans de telles situations, certaines sources sont employées au numérateur, d'autres au dénominateur, diminuant alors l'impact des discontinuités des unes et des autres sur l'indice observé. Mais l'utilisation conjointe de données d'origines variées n'est pas que le fait de ces ratios. Elle peut aussi avoir pour objet l'estimation d'une population, entre deux dates de recensement par exemple, et elle découle alors, si on utilise la méthode dite des composantes, de l'ajout des naissances et immigrants et de la soustraction des

Pays-Bas ou Angleterre (Monnier 1982). L'ajout aux nombres officiels espagnols des avortements de résidentes de ce pays ailleurs modifie complètement la perspective. Et encore, toutes les opérations clandestines sont omises!

Les erreurs de réponse sont de trois types: erreurs involontaires, volontaires ou non-réponses. Les premières tiennent plus au passage du temps qui engendre l'oubli et à l'importance, pour la personne interrogée, de l'objet de la question qu'à la mauvaise foi du sujet. Il arrive également qu'un individu doive répondre pour les autres membres du ménage qu'il peut connaître plus ou moins bien. Les erreurs volontaires sont plus subtiles. Elles peuvent découler de l'approche plus ou moins bien réussie du répondant, de la raison de la collecte, de l'opinion publique parfois exacerbée touchant l'information recherchée, ou encore des conditions sociales, économiques et politiques en vigueur au moment de l'interrogation. Un exemple intéressant de l'importance de ce dernier point est celui de la population albanaise de Yougoslavie qui s'est déclarée "turque" dans une bonne proportion au recensement de 1953 dans l'espoir de pouvoir émigrer en Turquie, ce pays n'accordant les documents officiels d'établissement sur son territoire qu'aux seuls ressortissants nationaux (Islami 1983). Quant aux non-réponses, elles proviennent d'une part de l'incapacité de la personne interrogée à fournir une réponse à certaines questions et d'autre part, des refus de réponse. De sérieux problèmes peuvent être engendrés par la méconnaissance ou le manque d'opinion des gens sur un sujet: il ne savent pas et le taux de non-réponse tend à s'accroître avec la complexité de la question. Les refus de réponse, quant à eux, peuvent toucher soit une ou plusieurs questions, soit l'ensemble et trouvent leur origine dans le manque d'intérêt, le thème, les conditions socio-économiques environnantes, la raison de la collecte et le niveau de choix laissé à l'individu quant à son obligation de répondre ou non. Ces deux dernières causes sont bien illustrées dans l'exemple suivant. Le taux de réponse des Juifs aux recensements tunisiens n'était pas très élevé. En 1941, le gouvernement du pays décidait de lier le recensement au statut des Juifs et, pour s'assurer de leur participation, a fait planer des menaces de sanctions sévères contre les récalcitrants. La population juive dénombrée a donc été beaucoup plus nombreuse qu'habituellement et que ne le laissent présager les recensements encadrants de 1936 et 1946 (Taieb 1982). Les erreurs de réponse peuvent aussi provenir d'erreurs des énumérateurs, lorsqu'ils sont utilisés pour la collecte.

En effet, l'énumérateur, cet intermédiaire entre le répondant et le questionnaire et/ou formulaire à compléter, peut introduire des erreurs de réponse soit en omettant une ou plusieurs questions, soit en oubliant d'enregistrer ou en le faisant mal, la réponse fournie. Son intervention à toutes les étapes de la collecte, jointe à son manque d'expérience (en règle générale) et sa méconnaissance des concepts de base font que les erreurs, bien involontaires, qu'il engendre comptent pour la plus grande part de l'erreur totale (Kalbach et McVey 1971, p. 11). Outre les erreurs de réponse susmentionnées, l'énumérateur peut être responsable d'erreurs de classification, due à la confusion des définitions; d'erreurs de couverture, par oubli d'individus ciblés ou inclusion de personnes ne l'étant pas. Le non-emploi d'énumérateurs ou d'enquêteurs n'élimine toutefois pas ces dernières erreurs, en particulier l'erreur de couverture.

L'erreur de couverture est la différence entre le sous-dénombrement brut (proportion, parmi les personnes ciblées, de celles qui ont été omises) et le sur-dénombrement brut (proportion, parmi les personnes ciblées, de celles qui ont été comptées deux fois ou plus, à laquelle s'ajoute la proportion des personnes non ciblées mais tout de même énumérées). Ses causes, outre les erreurs dues aux énumérateurs, vont du lieu et de la saison (difficile de recenser au Yukon en hiver!), de la raison de la collecte et de la réaction des enquêtés devant cette intrusion de leur vie privée au degré d'obligation de réponse imparti et à la réputation de l'individu ou de l'organisme responsable. L'ensemble de la population visée peut être affectée ou uniquement une région ou une certaine catégorie d'individus. Le niveau de l'erreur de couverture n'est pas toujours connu, mais lorsqu'il l'est, il peut être et est parfois corrigé. Les discontinuités dans les séries chronologiques peuvent donc découler tant des variations du degré de couverture dans le temps, s'il n'y a pas correction, que de la décision, à une date donnée, de corriger alors qu'on ne le faisait pas avant ou que des modifications dans les techniques de mesure des taux de sur- et sous-dénombrement ou dans les méthodes de correction. L'étude des séries chronologiques touchant à la population australienne ou anglaise, par exemple, doit prendre en considération le fait que depuis une dizaine d'années environ les effectifs estimés sont majorés du sous-dénombrement net (sous-dénombrement brut moins sur-dénombrement brut).

Tout comme les erreurs de collecte, les erreurs de traitement trouvent leur origine dans de multiples comportements et opérations: les procédures de correction, d'ajustement et de pondération. Toutes trois peuvent engendrer des discontinuités sur le long terme si leur méthodologie est modifiée ou si l'objet de la procédure change. Les procédures de correction ne concernent que les erreurs de réponse. La procédure va de la décision de corriger ces erreurs à la méthode pour le faire (lissage des résultats, allocation de réponses au pro-rata ou de façon aléatoire parmi les autres formulaires de mêmes caractéristiques). La source de discontinuités peut donc être tant une série de données corrigées s'ajoutant à une série non corrigée, qu'un

(Monnier 1982). Quant aux enregistrements continus, leur totalisation se doit de toujours concerner le même intervalle de temps, en terme de nombre de mois et de type d'année (de calendrier, fiscale, censitaire, scolaire). Par exemple, en République Fédérale Allemande, alors que, depuis 1977, le compte des avortements est sur 12 mois, il ne concerne que six mois et quelques jours en 1976.

Une fois les données de base recueillies, elles sont traitées et une des étapes du traitement en est la codification. Celle-ci consiste à allouer à la réponse de chaque individu un code qui correspond à une catégorie dans une nomenclature établie soit par l'organisme responsable ou le chercheur lui-même, soit, plus souvent, par des organismes internationaux spécialistes de la question traitée. L'utilisation de ces classifications internationales a pour avantage certain une meilleure comparaison des structures nationales et d'ailleurs, est amplement préconisée. Cependant, quelle que soit la typologie choisie, elle risque d'être fréquemment remaniée dans le temps, au gré de l'évolution de la société et des connaissances. F. Meslé et J. Vallin (1981) se sont frottés à ce problème lors de l'étude des causes de décès sur le long terme en France (7e, 8e et 9e révisions de la classification internationale des maladies de l'O.M.S.). Les modifications peuvent être légères ou affecter le principe-même du classement.

Les classifications, regroupements, modifications ne se font pas qu'à la collecte et au traitement. Les caractéristiques retenues dans les tableaux sabrent encore dans l'information qui parviendra à l'utilisateur, à moins qu'il n'ait la possibilité technique et souvent financière d'accéder directement aux données traitées. En bref, soulignons que les renseignements déjà classifiés précédemment risquent d'être regroupés en plus vastes catégories lors de la publication, celles-ci pouvant varier bien sûr d'une année à l'autre. Enfin, toutes les données disponibles ne seront pas publiées mais certaines pourront être acquises sur simple demande. Les choix faits sont évidemment reconsidérés à chaque parution et il s'en suit, encore une fois, des possibilités de rupture.

Enfin, une dernière discontinuité n'affecte que les sources de données à enregistrement continu. Par définition, un registre de données se modifie constamment par l'ajout et le retrait d'événements affectant la population ou l'élément visés, et ce, en théorie, dès leur survenance. Malheureusement, et c'est là une lacune bien connue de ce type de source, les délais d'enregistrement sont parfois fort longs et obligent à une mise à jour des fichiers mensuels durant plusieurs mois. Mentionnons, par exemple, que les fichiers de récipiendaires éligibles des allocations familiales un mois donné sont mis à jour tous les six mois durant deux ans (fichiers M0023 et M0024). Il importe alors que soient utilisés, dans une étude chronologique basée sur de tels registres, les fichiers sinon définitifs du moins à un même niveau de complétude (1e, 2e, 3e révisions). Il y va de leur comparabilité en terme de mise à jour et de qualité. Et en ce sens, cette discontinuité potentielle pourrait presque en être une de deuxième niveau.

2.1.2 Discontinuités de deuxième niveau

Les discontinuités de deuxième niveau tiennent essentiellement à la qualité de l'information recueillie ou disponible. Quels que soient le souci du détail lors de la planification, la collecte, le traitement, l'expérience et l'efficacité du personnel impliqué, l'ampleur des moyens mis en oeuvre, il est quasiment impossible d'en arriver à une collecte parfaite de données, répondant à tous les critères imaginables de qualité. Et ceux-ci sont nombreux. La qualité globale ou l'erreur totale d'une source de données particulière est tout simplement la somme des diverses erreurs, pour la seconde; la différence à "1" de cette somme pour la première.

La première disparité entre l'élément réel et mesuré découle de l'erreur dans les concepts, i.e. du degré de pertinence des questions posées. Comme montré à la section précédente, plusieurs questions différentes pourraient être retenues pour connaître un même état, chacune d'entre elles présentant des résultats plus ou moins concordants.

Il est généralement admis que la plus grande part de l'erreur totale provient, cependant, des erreurs de collecte, qu'il s'agisse de l'adéquation de la méthode de collecte choisie à l'information recherchée, des erreurs de réponse, de l'énumérateur ou de couverture.

Diverses méthodes de collecte d'informations peuvent être employées pour cibler un même objet d'étude auprès d'une même population. Toutes, cependant, ne sont pas aussi bien adaptées et ne sauront produire les mêmes résultats. Posons le cas suivant: nous désirons connaître le nombre total d'avortements subis une année donnée par les Espagnoles. Sachant que l'avortement est légalisé, sous certaines conditions, depuis 1985 et que chaque opération amène la complétion d'un formulaire statistique, nous pourrions ne nous fier qu'aux chiffres officiels. Or, la loi sur l'avortement a suscité une vive opposition en Espagne et un nombre important de médecins invoquent la clause de conscience pour se soustraire à sa pratique. Il s'en suit que le nombre officiel d'interruptions de grossesse est très faible (Monnier 1986). Mais devant ces difficultés, les Espagnoles vont se faire avorter dans un pays où la pratique est plus libérale, La discontinuité la plus visible est certes celle qui découle d'un arrêt brusque de la cueillette d'une information habituellement collectée de façon continue, pour cause de guerre, famine, épidémie, révolution ou, simplement, coupures budgétaires. Il peut également s'agir de pertes ou destructions de documents, comme on le vit souvent en études historiques. Les lacunes sont d'autant plus importantes dans les cas de graves événements conjoncturels que, généralement, les sources de données d'appoint éventuelles sont également manquantes et que tout porte à croire que ces événements affectent l'élément ou le phénomène étudiés, sinon la population ciblée. Par exemple, l'analyse de l'évolution historique du nombre de patients en instituts psychiatriques en France doit s'interrompre de 1914 à 1919, la première guerre mondiale ayant empêché la collecte habituelle d'information (Meslé et Vallin 1981). Ces coupures sont, cependant, généralement, de courte durée, les services normaux reprenant dès la fin des crises.

D'origine différente mais d'effet identique est la disparition d'une question posée depuis un certain temps. Une telle disparition peut n'être que temporaire mais généralement elle est permanente puisqu'elle découle d'une perte d'intérêt sociale ou économique du concept. Un exemple récent est le retrait en 1978 dans les publications relatives à l'état civil, du tableau portant sur la proportion d'accouchements à l'hôpital, au Canada. Ce tableau n'avait plus sa raison d'être, d'une part parce que le problème de base, la mortalité maternelle et infantile, est à toutes fins pratiques négligeable, d'autre part, parce que au cours des dix dernières années de cueillette, les proportions étaient quasi-stationnaires à presque 100%. Peut-être, cependant, avec la montée des méthodes naturelles d'accouchement à la maison, assisterons-nous, dans quelques années, à la réintroduction de l'information? Une autre cause à la base du retrait d'une question peut être la pression sociale de l'opinion publique. Les responsables voudraient bien poser la question, les chercheurs en connaître la réponse mais le thème ciblé exacerbe la population. Il en a été ainsi autrefois du divorce, puis de l'union libre et l'avortement. Ça l'est parfois des questions ethniques ou raciales. Dans de tels cas, la question peut être posée puis omise plusieurs fois au gré des fluctuations de l'opinion publique.

Un peu moins apparente mais tout aussi problématique est la modification d'une question. De prime abord, les données sont disponibles régulièrement dans le temps. Cependant, ce qu'elles couvrent a varié. La discontinuité est plus ou moins importante selon l'objet de la modification. Il peut s'agir du concept sous-jacent à un même terme, comme par exemple, en France, lorsqu'on parle de "rang des naissances légitimes" dans les statistiques de l'état civil. De 1949 à 1964, on sous-entendait par là "rang par femme" alors que depuis 1965, on parle de "rang dans le mariage présent", ce qui est tout à fait différent dans les cas de mariages multiples (J.-L. Rallu 1986). L'objet de la modification peut aussi être la formulation elle-même de la question. En effet, l'information désirée peut être obtenue de plusieurs façons et la question retenue n'est qu'une parmi diverses formulations possibles. L'âge en est un bon exemple. Il peut être obtenu directement ou encore via une question sur la date de naissance, l'une et l'autre solutions ne menant pas nécessairement à la même réponse, comme l'ont montré de nombreuses études. Enfin, certaines questions sélectives, n'ayant d'autre but que d'identifier le répondant recherché mais néanmoins utilisables en recherche, peuvent être modifiées sans qu'on y prête beaucoup attention, croyant le changement sans conséquence. Dans certains cas, pourtant, il s'avère fondamental. Généralement, et le concept et la formulation sont alors changés. L'exemple typique en est la personne responsable de la complétion du recensement dans un ménage. Rôle autrefois dévolu au "chef de ménage", i.e. l'homme dans un couple, il peut maintenant l'être à la femme, dûment identifiée alors comme "personne 1" au Canada, "personne de référence" en France.

Les discontinuités ne proviennent pas que des questions mais touchent aussi à la population ciblée. Celle-ci peut être élargie ou restreinte pour différents motifs. Le premier est le changement des limites territoriales, politique (adjonction ou cession de nouveaux territoires pour un pays) ou simplement statistique (révision des frontières des régions statistiques). Le deuxième motif de changement de la population ciblée en est simplement un de définition. En gros, les organismes statistiques s'entendent pour proposer deux types de population: "de fait", i.e. sur place lors de la collecte d'information, ou "de jure", i.e. habituellement résidente au lieu recensé. Le passage de l'un à l'autre dans le temps peut provoquer des écarts majeurs dans les données. L'exemple en est le Portugal, "de fait" avant 1940, "de jure" par la suite (Monnier 1982). D'autres pays, tout en continuant de cibler le même type de population, changent le concept sous-jacent. C'est en particulier le cas de la France, qui a vu sa population de droit (de jure) recensée s'enrichir depuis 1962 des militaires stationnés hors métropole (Monnier 1982). La définition des populations ciblées par d'autres sources que le recensement peut aussi être modifiée, tel que le montre la statistique française des aliénés qui n'inclut plus, depuis 1949, que les malades "des établissements publics ou privés faisant fonction d'établissements publics" négligeant les hôpitaux privés antérieurement visés (Meslé et Vallin 1981).

Une autre source de discontinuités sur le long terme est la périodicité de la collecte, effectuée sporadiquement, ou de façon continue. La collecte sporadique pour être comparable doit avoir lieu à intervalles réguliers et donc à date fixe, ce qui est loin d'être toujours le cas. Les exemples abondent sur ce point, les dates des recensements nationaux successifs étant souvent différentes encore, ne connaît-on pas le niveau de comparabilité des différentes questions sur un même thème dans le temps, des modifications pouvant survenir à tellement de niveaux qu'il devient alors malaisé d'en évaluer la résultante en terme de qualité et de pertinence. Le chercheur fait donc face à ce que nous allons nommer des "discontinuités dans les séries chronologiques".

Un autre point reste à souligner. L'analyse de séries chronologiques ne se base pas que sur l'information relevée sur le terrain, celle-ci ne l'étant, généralement, qu'à intervalles de quelques années. Il est d'usage, pour les années intermédiaires, de faire appel à des estimations que le chercheur produit lui-même ou encore qu'il emprunte aux travaux de ses confrères ou aux publications statistiques officielles. Dans ce dernier cas, il est soumis à l'éventualité d'une modification, au cours des ans, de la méthodologie d'estimation ou encore de l'information de base utilisée.

Ce bref tour d'horizon n'est guère encourageant. Qui serait alors assez téméraire pour décider d'une analyse sur le long terme, sachant que les séries qu'il utilise sont possiblement truffées de pièges pouvant l'induire en erreur, soient en elles-mêmes soit dans le calcul d'indices quelconques? Comment procéder? De prime abord, le chercheur doit identifier les faiblesses de ses données et tenter de les corriger. Pour l'aider à y parvenir, nous avons tenté de dresser, dans la présente étude,une liste, la plus exhaustive possible, des sources de discontinuités éventuelles, illustrant lorsque faire se peut chaque cas par un exemple concret. Elles sont nombreuses et, par souci de clarté, ont été classées par types. Puisqu'elles posent problème dans l'analyse, nous avons ensuite dégagé les comportements possibles pour leur faire face et le cas échéant les corriger. Enfin à partir d'un exemple concret, nous allons tenter de mesurer l'impact de différentes discontinuités sur quelques indicateurs démographiques et de déduire des résultats observés un schème de fonctionnement des discontinuités dans les séries chronologiques.

2. CAUSES DE DISCONTINUITÉS

Les discontinuités dans les séries chronologiques jouent à deux niveaux entre le réel que l'on voudrait bien étudier et l'observation où se situent en fait les données malheureusement imparfaites dont on dispose. Les discontinuités de premier niveau sont apparentes et concernent directement les faits observés: elles posent problème au premier abord. Les discontinuités de deuxième niveau sont plus subtiles, en retrait, sous-jacentes, tellement qu'on les omet généralement lorsqu'on parle de "discontinuités". Elles sont derrière les chiffres publiés et s'attachent en fait à leur qualité, touchant à l'écart dans le temps entre valeurs réelles et observées. Les discontinuités de premier niveau concernent donc les formats des éléments étudiés tandis que celles du deuxième biaisent plutôt les tendances.

Les discontinuités des deux niveaux affectent les séries chronologiques avec plus ou moins d'intensité et de complexité selon d'une part, le nombre de sources utilisées dans le temps et d'autre part, leur emploi simultané ou non. Notre présentation des discontinuités se fera donc en quatre temps, en fonction des deux critères précédents, et par niveau croissant de complexité: (1) changements dans une source dans le temps; (2) changement de sources dans le temps chronologiquement; (3) changement dans les relations entre sources utilisées conjointement, dans le temps; (4) changement des sources utilisées conjointement, chronologiquement.

Pour chaque catégorie, les discontinuités présentées seront scindées par niveaux. Elles sont nombreuses mais nous désirons nous excuser à l'avance des omissions toujours possibles. Elles concernent plusieurs types de sources: enquêtes et sondages; registres de données; fichiers administratifs; recensements; études, travaux, calculs statistiques d'autres auteurs et/ou organismes. Certaines de ces sources font partie d'un programme d'enregistrement continu (registres de données, par exemple l'état civil; certains fichiers administratifs, tels ceux des allocations familiales de Santé et Bien-Être social Canada; d'assurance-maladie provinciaux; d'immigrants d'Emploi et Immigration Canada) ou sporadique (le recensement; certaines enquêtes comme par exemple, l'enquête sur la population active canadienne) et peuvent, donc, à elles seules, servir de base à une analyse chronologique. Les discontinuités qui peuvent les toucher dans le temps seront étudiées en premier lieu.

2.1 Changements dans une source dans le temps

Le souci constant d'amélioration d'un système de collecte, qu'il soit gouvernemental ou privé, conduit nécessairement à des coupures périodiques dans les séries historiques, coupures soit apparentes (discontinuités de premier niveau) soit cachées (de deuxième niveau).

2.1.1 Discontinuités de premier niveau

Les causes des discontinuités de premier niveau se retrouvent à toutes les étapes du processus d'obtention de l'information, de la préparation des questionnaires, au traitement et à la publication des données. Elles peuvent même y être extérieures, comme nous le verrons plus loin.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

DISCONTINUITÉS DANS LES SÉRIES CHRONOLOGIQUES

Céline Fortier[1]

RÉSUMÉ

Cet article présente la liste exhaustive des discontinuités qui peuvent affecter les séries chronologiques de données, par type et niveau de complexité. Elles sont nombreuses et ne peuvent être négligées. Tout chercheur se doit d'identifier celles qui touchent sa série particulière et de prendre une décision quant à la façon de poursuivre son étude (abandon, changements dans les spécifications ou corrections). Le meilleur outil pour y parvenir est le test de sensibilité qui permet de mesurer l'impact de chaque discontinuité relevée. Le niveau de cet impact est variable et l'analyse de l'évolution de sept séries de mesures de l'indice synthétique de fécondité au Canada de 1950 à 1986 permet de dégager certains postulats quant à leur variation. Quoiqu'il en soit, il importe de ne pas mettre l'emphase sur des différences minimes dans le temps mais plutôt sur les variations majeures et les tendances soutenues.

MOTS CLÉS: Discontinuités; test de sensibilité; indice synthétique de fécondité; correction de données; analyse de données.

1. INTRODUCTION

À la base de toute analyse de séries chronologiques, il y a le chercheur, une idée à étudier et généralement, une ou plusieurs hypothèses à tester. La première étape demeure l'obtention de données sur le long terme, comparables dans le temps, pertinentes et le moins éloignées possible de l'objet de l'étude tel qu'il est réellement. Or, à moins qu'il puisse créer sa propre banque de données (via une enquête rétrospective elle-même sujette à des discontinuités), le chercheur fait face à des contraintes et limites d'ordre technique: il est lié aux données disponibles. Il devra juxtaposer les résultats de différents recensements, enquêtes et/ou sondages.

Or, toute collecte d'informations est faite dans un but précis, pour apporter réponse à des questions d'importance, pendantes au présent, selon les besoins des utilisateurs. Elle est soumise à des contraintes telles la susceptibilité ou la capacité à répondre de façon exacte des personnes interrogées, l'opinion publique ou encore le développement économique et social, le milieu politique ou les impératifs économiques de l'individu ou de l'organisme responsables. L'information collectée doit ensuite être traitée et ce traitement dépend d'une part, des questions retenues et d'autre part, de l'infrastructure mise en place, équipement technique ou main-d'oeuvre, et de l'expertise non seulement disponible mais existante. Enfin, les données résultantes sont compilées et publiées, généralement sous forme de tableaux, par catégories qui sont fonction, là encore, des demandes et besoins, sinon des typologies en vigueur.

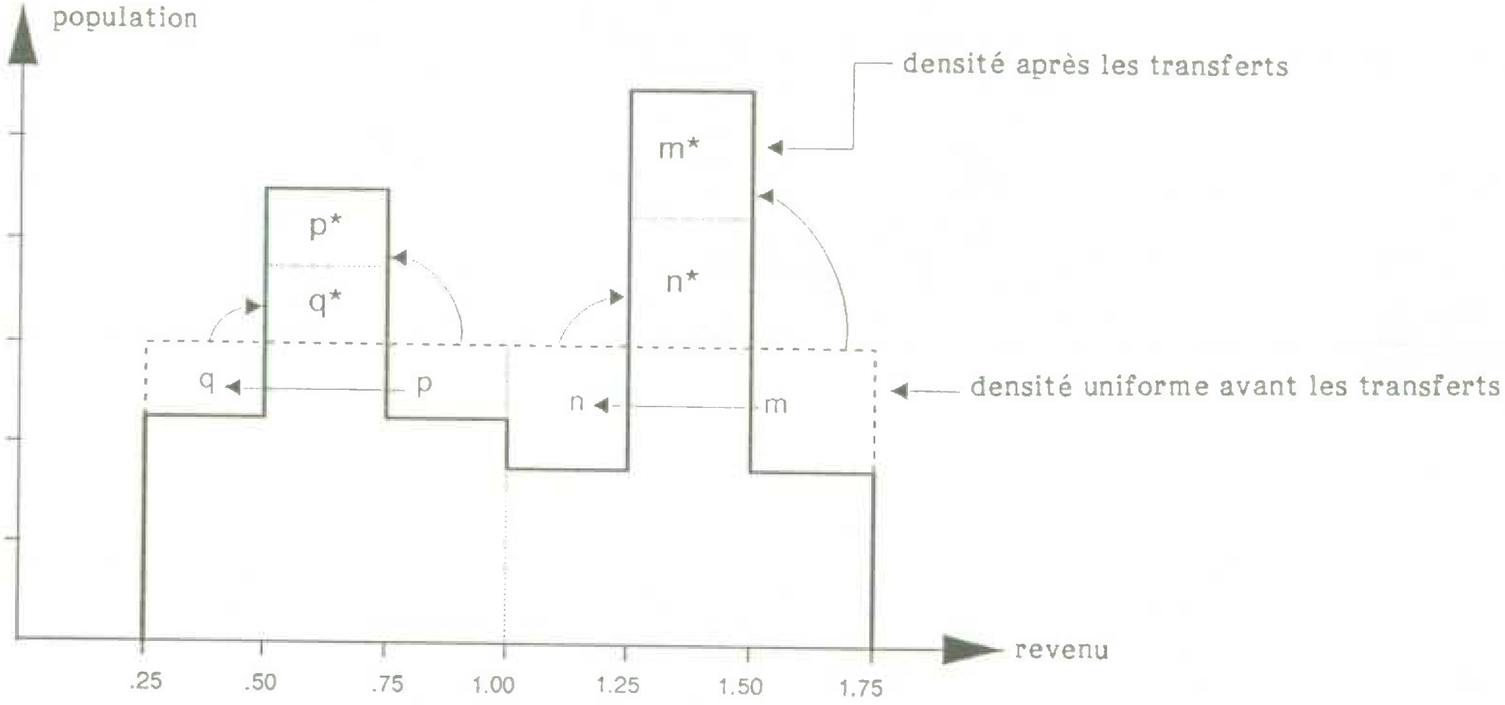
Seulement, le temps passant, les préoccupations changent, les besoins évoluent tout comme les mentalités et ce qui était tabou hier peut être mode aujourd'hui. Le niveau d'instruction moyen croît avec le développement socio-économique et s'en suit l'augmentation des niveaux de vie, le changement des styles de vie. Les avancées technologiques et méthodologiques accroissent la rapidité et la qualité du traitement. De plus, l'opinion publique, conscientisée à de nouveaux problèmes, tels le jumelage des fichiers ou la confidentialité des données et le respect de la vie privée, impose aux collecteurs d'informations une nouvelle problématique qu'ils ne peuvent négliger. Enfin, en bout de ligne, les publications changent en nombre et en contenu, en réponse à la collecte et au traitement, bien sûr, mais aussi aux besoins des utilisateurs, à la diversification des intérêts et au raffinement croissant des classifications en usage.

Il s'en suit que, d'une collecte à l'autre, l'information demandée et publiée se modifie. Bien que très au fait des exigences de la recherche et tenant compte du souci de comparabilité des données dans le temps, les responsables se doivent de retirer des questions, soit temporairement, soit définitivement, lorsqu'elles ne répondent plus aux attentes et ne présentent aucun intérêt économique, social ou politique. Ils se doivent également de modifier la formulation de certaines autres, soit pour les réactualiser soit par suite de l'évolution de l'acceptabilité sociale de l'élément visé. Enfin, régulièrement, ils en soumettent sur de nouveaux thèmes, le nombre total de.questions retenues étant cependant fonction de considérations budgétaires.

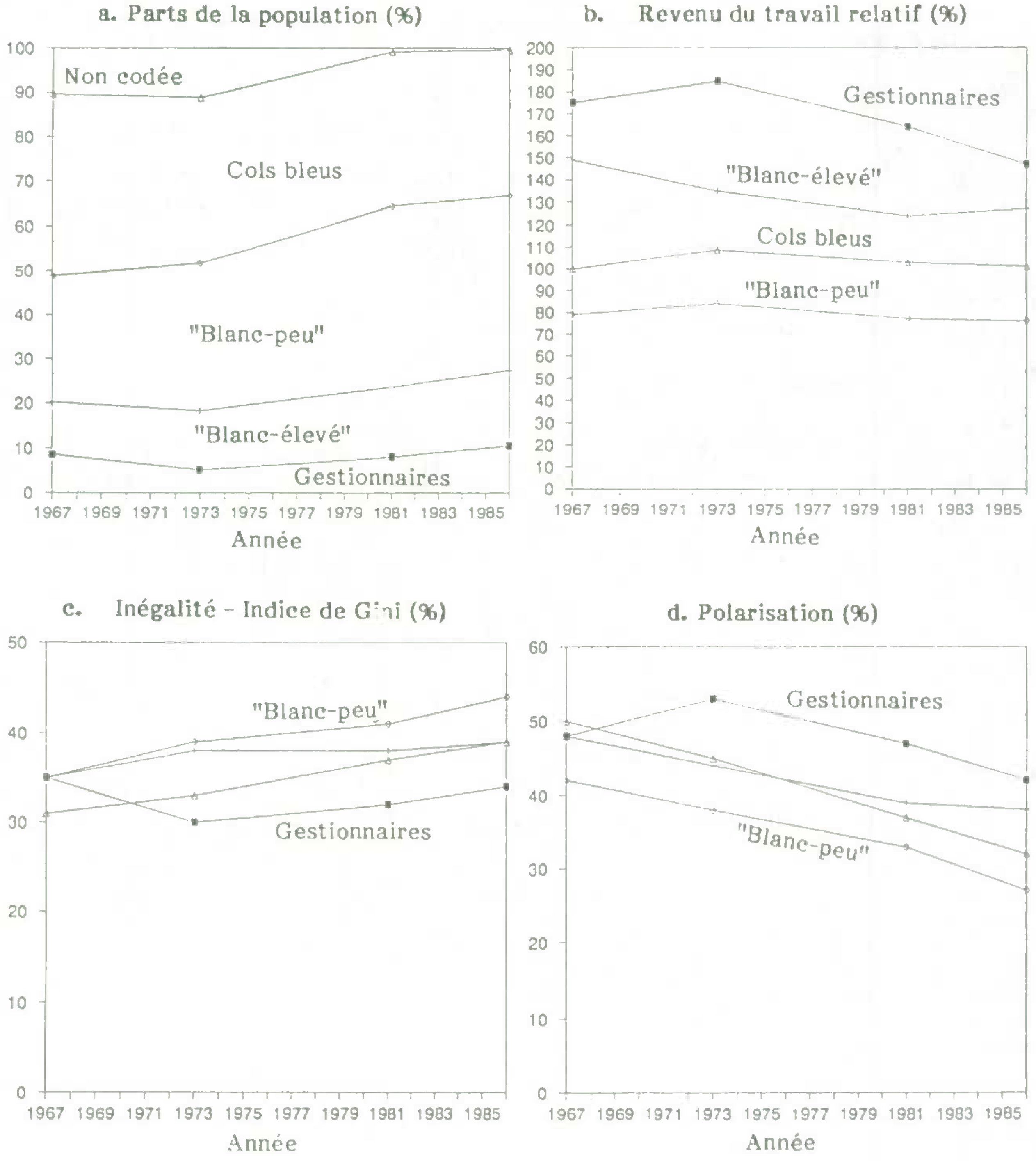
Donc, non seulement, n'y a-t-il aucune certitude qu'à une date donnée, la question posée soit la plus pertinente à cerner le problème identifié, puisqu'elle résulte de compromis divers, mais

[1] Céline Fortier, Division de la démographie, Statistique Canada, Ottawa, Ontario K1A 0T6.

Figure 1: Illustration de la polarisation et de l'inégalité


population
densité après les transferts
m*
p*
n*
q*
q
p
n
m
densité uniforme avant les transferts
revenu
.25
.50
.75
1.00
1.25
1.50
1.75

Graphique 5 — Tendances selon la profession
Particuliers PVPA, âgés de 15 ans et plus, revenu du travail



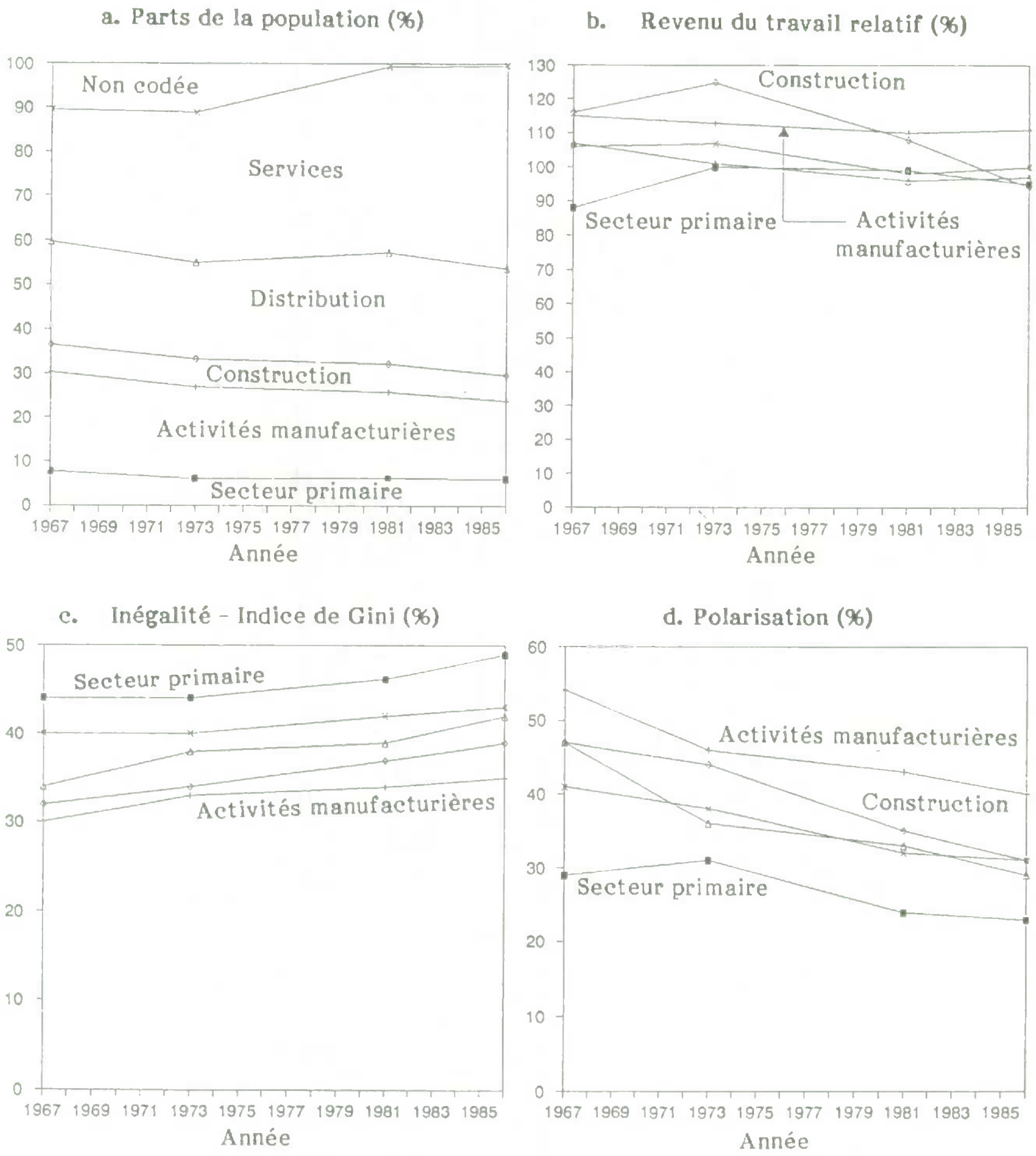
a. Parts de la population (%)

b. Revenu du travail relatif (%)

c. Inégalité - Indice de Gini (%)

d. Polarisation (%)

Graphique 4 -- Tendances selon la branche d'activité
Particuliers PVPA, âgés de 15 ans et plus, revenu du travail



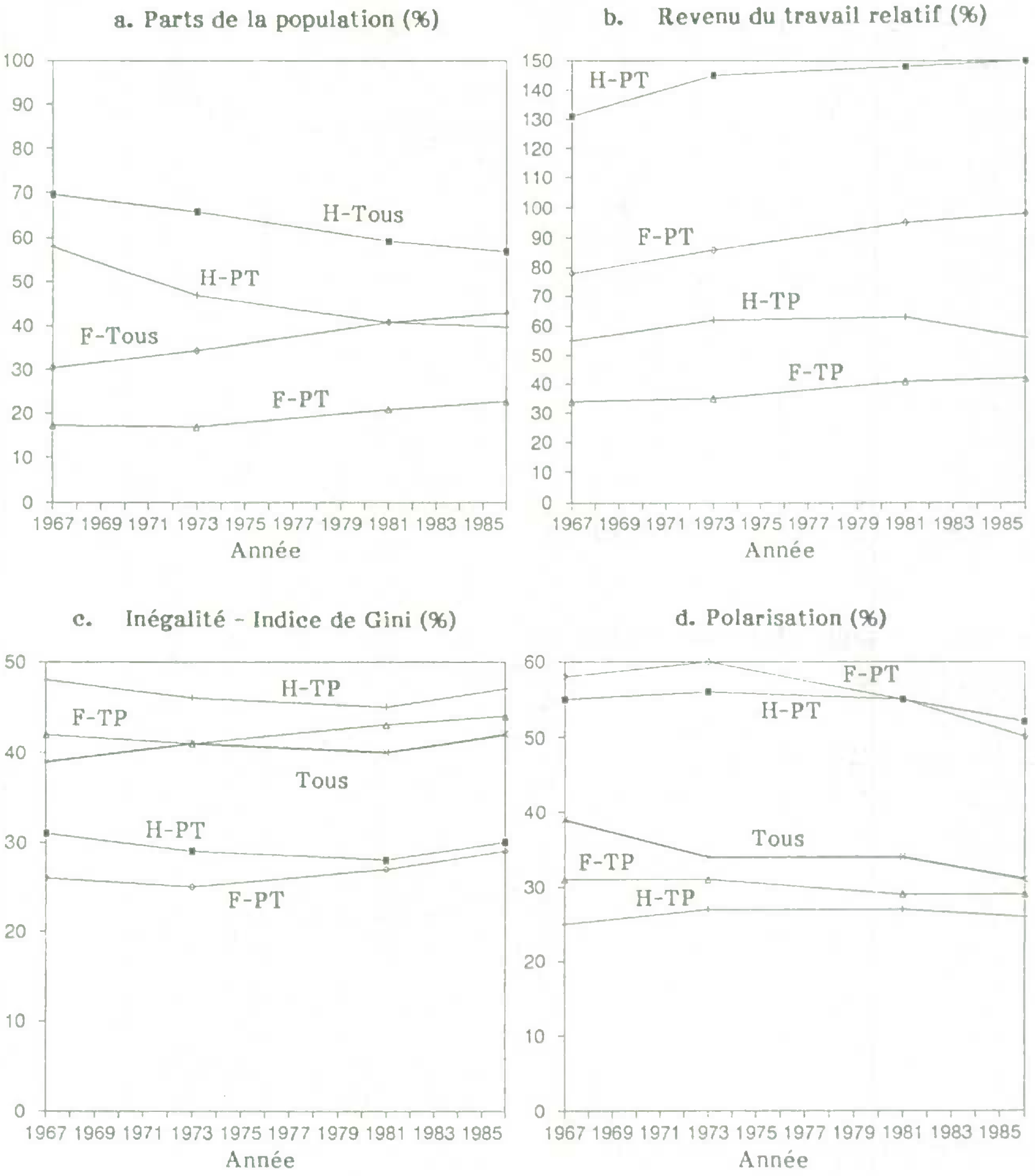
a. Parts de la population (%)

b. Revenu du travail relatif (%)

c. Inégalité - Indice de Gini (%)

d. Polarisation (%)

Graphique 3 — Tendances selon le sexe et le statut de travailleur à plein temps ou à temps partiel
Particuliers PVPA, âgés de 15 ans et plus, revenu du travail



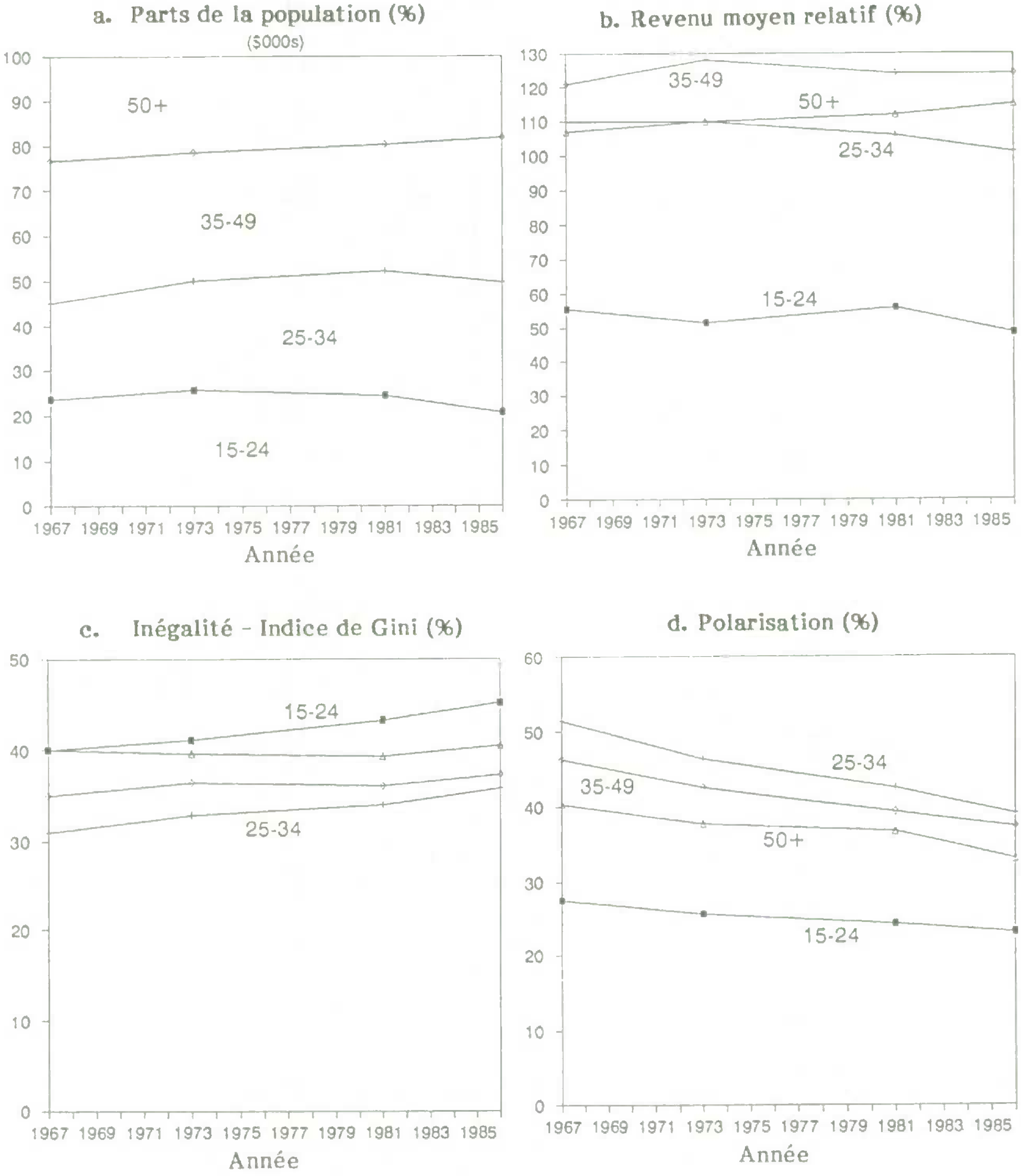
a. Parts de la population (%)

b. Revenu du travail relatif (%)

c. Inégalité - Indice de Gini (%)

d. Polarisation (%)

Graphique2 -- Tendances selon l'âge
Particuliers PVPA, âgés de 15 ans et plus, revenu du travail



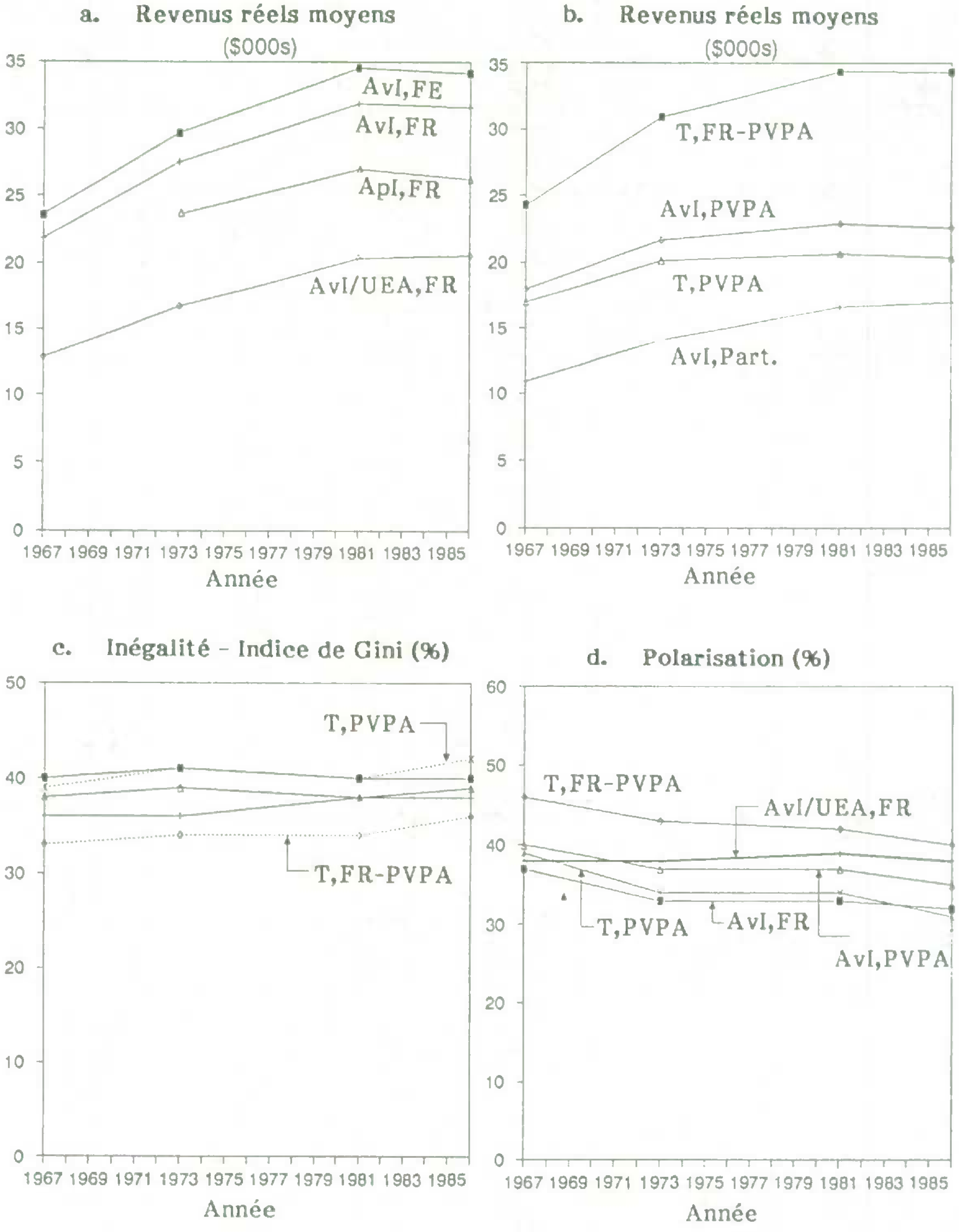
a. Parts de la population (%)

b. Revenu moyen relatif (%)

c. Inégalité - Indice de Gini (%)

d. Polarisation (%)

Graphique 1
Tendances par unité déclarante et par concept du revenu



critère de Lorenz, elles sont construites pour répondre de façon cohérente aux mouvements d'une fonction de répartition tendant vers la bimodalité. Une telle mesure, WPOL, basée sur l'indice de Gini, est élaborée dans Wolfson (1986a) et elle a été utilisée dans le document du Conseil économique (1987).

La mesure WPOL n'a pas été employée ici parce qu'on l'a jugée plus complexe et moins compréhensible que cela n'était nécessaire pour l'analyse. La mesure de la polarisation qui a été utilisée surtout, la part de la population dont les revenus sont compris entre 75% et 150% de la médiane, bien qu'elle ne soit pas nécessairement compatible de façon formelle avec le concept de polarisation, est facile à comprendre et on en a vérifié la compatibilité chaque fois qu'on l'a employée. La mesure WPOL est compatible de façon formelle avec le concept de polarisation et on aurait pu l'utiliser à la place de la mesure employée. Cependant, les résultats généraux auraient été les mêmes et il aurait fallu que le lecteur se familiarise avec une nouvelle statistique complexe avant de se sentir à l'aise avec les résultats. Si le concept de la classe moyenne qui recule continue quand même de présenter de plus en plus d'intérêt, il pourrait se révéler avantageux d'élaborer un ensemble plus formalisé de mesures statistiques telle que la mesure WPOL par des besoins analytiques.

Tableau 1
Statistiques sur l'inégalité et la polarisation, toutes les familles de recensement, revenu total

	Année			
	1967	1973	1981	1986
Revenu moyen ($ de 1986)	21,850	27,550	31,900	31,650
Revenu médian ($ de 1986)	19,450	23,900	27,500	26,200
Parts des quintiles (%)				
0 - 20	3.5	3.2	4.1	4.3
20 - 40	10.7	9.9	10.3	9.9
40 - 60	17.8	17.3	17.3	16.7
60 - 80	24.9	25.5	25.5	25.0
80 - 100	43.0	44.0	42.9	44.1
Indice de Gini (%)	39.8	41.3	39.5	40.4
Parts de la population par gamme de revenu médian				
< 50%	25.1	26.5	25.2	24.9
50 - 75	12.1	12.3	12.9	13.4
75 - 125	26.1	22.5	23.1	22.3
125 - 150	11.1	10.2	10.3	9.4
> 150%	25.6	28.6	28.6	30.0
75 - 150	37.2	32.7	33.4	31.7

Tableau 2
Statistiques sur l'inégalité et la polarisation, particuliers PVPA, âgés de 15 ans et plus, revenu du travail

	Année			
	1967	1973	1981	1986
Revenu moyen ($ de 1986)	16,950	20,150	20,700	20,400
Revenu médian ($ de 1986)	15,050	17,250	18.050	17,400
Parts des quintiles (%)				
0 - 20	3.9	3.6	3.6	3.4
20 - 40	11.0	10.2	10.2	9.5
40 - 60	17.9	17.2	17.4	17.0
60 - 80	24.6	25.2	25.7	25.7
80 - 100	42.5	43.8	43.1	44.4
Indice de Gini (%)	38.9	40.7	40.2	41.8
Parts de la population par gamme de revenu médian				
< 50%	24.1	25.4	26.1	27.2
50 - 75	12.3	11.8	12.1	12.2
75 - 125	26.8	23.7	23.4	21.5
125 - 150	12.5	10.6	10.2	9.3
> 150%	24.4	28.5	28.1	29.8
75 - 150	39.3	34.3	33.6	30.8

ne devrait jamais utiliser de telles statistiques dans ces contextes parce qu'elles ne mesurent tout simplement pas ce qu'elles prétendent mesurer.

Cependant, le fait que les mesures de l'inégalité, même si elles sont parfaitement compatibles avec le critère de Lorenz, puissent être incompatibles avec le concept désiré constitue un problème plus fondamental dans le contexte des discussions portant sur la classe moyenne qui recule. Cette situation est illustrée à la figure 1, qui montre deux densités de fonction de répartition du revenu hypothétiques simples. La première est une fonction de répartition uniforme sur la gamme de revenus allant de 0.25 à 1.75, elle est illustrée par une ligne discontinue.

La seconde densité, illustrée par une ligne continue, est clairement bimodale et sa partie du milieu est un peu affaissée. Nous soutiendrions que, selon toute définition raisonnable de la polarisation ou du recul du milieu, cette dernière densité est la plus polarisée.

Est-elle aussi plus inégale?

La réponse est non, sans équivoque. La seconde densité a été construite de telle façon que, selon toute mesure de l'inégalité qui est compatible avec le critère de Lorenz, elle est plus égale. Cela peut être démontré par le simple fait que l'on peut obtenir la fonction de répartition bimodale à partir de la fonction de répartition uniforme à l'aide de deux ensembles de transferts de redistribution progressifs qui préservent la moyenne, au sens donné par Atkinson (1970).

Le premier ensemble de transferts de revenu égalisateurs se fait à partir de particuliers de la partie p dans la partie de 0.75 à 1.00 de la gamme de revenus vers des particuliers de la partie q dans la partie la moins élevée, de 0.25 à 0.50. Les p donnent aux q des parties de leurs revenus égales à la moitié de la différence entre leurs revenus -- 0.25 en moyenne, de sorte qu'ils se déplacent aux parties p^* et q^* de la fonction de répartition bimodale, dans la gamme de revenus de 0.50 à 0.75. De même, les personnes dans la partie la plus élevée de la répartition des revenus, avec des revenus compris entre 1.50 et 1.75, partie m, donnent une moyenne de 0.25 de leur revenu aux personnes du groupe n, dans la partie moyenne supérieure de la fonction de répartition, celle des revenus allant de 1.00 à 1.25. A la suite de cet ensemble de transferts progressifs, ces deux groupes de personnes se trouvent dans la même gamme de revenus, de 1.25 à 1.50, dans les parties m^* et n^* de la fonction de répartition bimodale.

Ainsi, par construction, la fonction de répartition bimodale est en même temps plus polarisée et plus égale que la fonction de répartition uniforme d'où elle provient. La polarisation et l'inégalité sont donc manifestement des concepts différents, comme cela a été signalé pour la première fois dans Love et Wolfson (1976).

Ce résultat laisse subsister la question de savoir quelles statistiques devraient être utilisées pour mesurer la polarisation. Dans les documents portant sur la partie du milieu qui recule, certains auteurs utilisent, en plus des mesures de l'inégalité, les parts du revenu par quintiles, alors que d'autres ont employé la fraction de la population dans diverses gammes de revenus définies en fonction du revenu moyen ou médian. En fait, la figure 1 a été construite de façon particulièrement mauvaise pour ces genres de statistiques.

Puisque la fonction de répartition est symétrique, la moyenne est égale à la médiane qui est un. On peut démontrer que la part du tiers du milieu de la fonction de répartition bimodale est inférieure à la part du tiers du milieu de la fonction de répartition uniforme, alors que la part des deux tiers du milieu augmente au cours de la transition vers la fonction de répartition bimodale. Ainsi, les parts du revenu de divers groupes de quantiles du milieu **ne** sont **pas** nécessairement compatibles avec toute formalisation raisonnable du concept de polarisation. Cela signifie, à son tour, que le grand nombre d'articles prétendant analyser le recul de la classe moyenne qui utilisent des indicateurs de l'inégalité comme les parts de quintiles (p. ex., Levy 1987; Beach, 1988) sont tout simplement incapables de détecter le phénomène qu'ils prétendent étudier.

De plus, la part de la population avec des "revenus de niveau moyen" augmente ou diminue selon la façon dont le mot "moyen" est défini dans cet exemple. On voit cela facilement en inspectant la figure 1. La population dont les revenus se trouvent à moins de 25% de la moyenne ou de la médiane diminue manifestement, mais la population dont les revenus se trouvent à moins de 50% de la moyenne ou de la médiane augmente. Ainsi, les statistiques qui comptent la part de la population avec des revenus "près de la moyenne" **ne** sont **pas** nécessairement compatibles, elles non plus, avec une définition raisonnable de la polarisation.

Tout n'est pas perdu, cependant. Compte tenu de notre compréhension explicite améliorée du concept de polarisation basée sur l'analyse de la figure 1, nous pouvons choisir un ensemble de statistiques pour faire une analyse détaillée. Les parts de la population dans diverses gammes de revenus définies en fonction du revenu médian, comme dans les tableaux 1 et 2 de l'exposé principal, constituent un bon exemple. On peut alors utiliser un indicateur statistique sommaire comme la part de la population dont les revenus sont compris entre 75% et 150% de la médiane, à des fins de discussion et de présentation graphique, pourvu que l'analyste le vérifie toujours afin de s'assurer qu'il est compatible avec les données plus détaillées. C'est ce qui a été fait pour l'analyse mentionnée dans l'exposé principal.

De plus, il y a des approches formelles plus complexes. Elles définissent une classe d'indicateurs sommaires de la polarisation qui sont analogues aux mesures de l'inégalité sauf que plutôt que d'être compatibles avec le

Harrison, B., C. Tilly, and B. Bluestone (1986), "Wage Inequality Takes a Great U-Turn", *Challenge*, March-April.

Kuttner, B. (1983), "The Declining Middle", *The Atlantic Monthly*, July.

Leckie, N. (1988), *The Declining Middle and Technological Changes: Trends in the Distribution of Employment Income in Canada, 1971-84*, Discussion Paper No. 342, Economic Council of Canada, Ottawa, January.

Levy, F. (1987), "Changes in the Distribution of American Family Incomes, 1947 to 1984", *Science*, Vol. 236, pp. 923-27, May.

Love, R. and M.C. Wolfson (1976), *Income Inequality: Statistical Methodology and Canadian Illustrations*, Catalogue 13-559 Occasional, Statistics Canada, Ottawa, March.

Loveman, G.W. and C. Tilly (1988), "Good Jobs or Bad Jobs: What Does the Evidence Say", *New England Economic Review*, January/February.

Myles, J. (1988) "The Expanding Middle: Some Canadian Evidence on the Deskilling Debate", *The Canadian Review of Sociology and Anthropoly*, Vol. 25:3, pp. 35-364, August.

Picot, G., J. Myles and Ted Wannell (1990), *Good Jobs/Bads Jobs and the Declining Middle 1967-1986*, Analytical Studies Branch Research Paper Series No. 28, Statistics Canada, Ottawa.

Rosenthal, N.H. (1985), "The Shrinking middle class: myth or reality?" *Monthly Labour Review*, March.

Wolfson, M.C. (1986a), "Stasis Amid Change -- Income Inequality in Canada 1965-1983", *Review of Income and Wealth*, December.

Wolfson, M.C. (1986b), "Polarization, Inequality, and the Disappearing Middle*", Statistics Canada, Ottawa, Mimeo.

Wolfson, M.C. and J. Evans (1989), *Statistics Canada's Low Income Cut-Offs — Methodological Concerns and Possibilities: A Discussion Paper*, Statistics Canada, Ottawa, December.

ANNEXE

La mesure de l'inégalité et de la polarisation

Les formalisations les plus généralement acceptées du concept d'inégalité économique sont toutes reliées à la courbe de Lorenz. Cette courbe constitue une façon de montrer toute répartition du revenu -- un ensemble de données qui montrent combien il y avait de personnes recevant un revenu (qu'il s'agisse d'unités familiales ou de particuliers) à divers niveaux de revenu. La courbe de Lorenz est un graphique qui montre la fraction cumulative de la population le long de l'axe horizontal et leur part cumulative du revenu (ou d'une autre mesure de la position économique) le long de l'axe vertical, en supposant que la population a été classée en ordre croissant de revenus.

Le classement de deux répartitions du revenu (p. ex., pour deux moments dans le temps) selon que la courbe de Lorenz d'une répartition du revenu se trouve à au moins un endroit au dessus et jamais au dessous de l'autre est le critère principal de presque toutes les bases axiomatisées de mesures de l'inégalité (Atkinson, 1970; Love et Wolfson, 1976; Cowell, 1977). La mesure, utilisée le plus souvent, de l'inégalité du revenu (ou d'un autre indicateur de la position économique) est l'indice de Gini. Il est entièrement compatible avec le classement des répartitions du revenu donné par les courbes de Lorenz.

Cependant, l'indice de Gini n'est pas le seul indice sommaire de l'inégalité qui est entièrement compatible avec les classements donnés par les courbes de Lorenz. Les autres indices comprennent les mesures de Theil, de Theil-Bernouilli, d'Atkinson et la mesure exponentielle ainsi que le coefficient de variation. Quand le classement selon les courbes de Lorenz est ambigu -- c.-à-d. quand les courbes de Lorenz pour deux répartitions du revenu se coupent de sorte qu'aucune des deux ne domine clairement l'autre, ces mesures donneront généralement des classements différents.

D'autres statistiques populaires de l'inégalité sont basées sur les quantiles, comme les parts du revenu qui reviennent aux quintiles de la population. Il est préférable de considérer les parts du revenu par quintile comme des indicateurs de l'inégalité plutôt que comme des mesures de l'inégalité. Bien qu'elles ne soient jamais incompatibles avec les classements selon les courbes de Lorenz, il se peut que les parts par quintile ne changent pas bien que les classements selon les courbes de Lorenz changent.

Il y a aussi des statistiques qui sont utilisées dans des discussions portant sur l'inégalité bien qu'elles ne soient même pas compatibles avec les classements selon les courbes de Lorenz. Ces statistiques comprennent la variance des logarithmes du revenu (p. ex., utilisée dans Harrison et coll., 1986) et le rapport interquartile. On et toutes montrent des tendances plus fortes à la baisse dans les proportions avec des revenus de niveau moyen. Ainsi, bien qu'il y ait eu des déplacements substantiels entre les secteurs de la population active de 1967 à 1986, particulièrement du secteur des activités manufacturières à celui des services, ces déplacements n'aident pas à expliquer l'augmentation de la polarisation. Cela est dû, tout simplement, au fait que la polarisation a augmenté substantiellement dans chacun des principaux secteurs industriels.

Profession: Finalement, les graphiques 5.a à 5.d montrent les tendances correspondantes pour les catégories professionnelles. Comme dans le cas de la classification par branches d'activité, la présente classification est très grossière et elle comprend un grand nombre de professions "non codées" en 1967 et en 1973. A nouveau, comme l'on pourrait s'y attendre, la diminution la plus importante a été relevée pour les métiers de cols bleus, alors que la plus grosse augmentation l'a été dans les emplois de cols blancs peu payés et exigeant peu de compétences ("blancs-peu" -- emplois de bureau, emplois dans la vente, emplois dans les services; "blancs-élevé" désigne les professions et les emplois techniques). A part des soubresauts dans les tendances pour le groupe relativement petit des gestionnaires, qui pourraient n'être rien d'autre qu'un artefact des changements dans le codage des professions de 1967 à 1973, les tendances relatives à l'inégalité et à la polarisation dans les catégories professionnelles sont généralement parallèles. Ainsi, bien que la composition par profession de la population active ait changé considérablement, on ne peut utiliser ce fait pour expliquer le "recul de la partie du milieu" de la population active canadienne, du moins pour les grosses catégories utilisées.

F. CONCLUSIONS

Beaucoup de documents ont été publiés aux Etats-Unis relativement à la "classe moyenne qui recule". Ces documents ont eu tendance à expliquer le phénomène en fonction de la "désindustrialisation", de la "déqualification", de l'érosion de métiers de cols bleus biens rémunérés dans le secteur des activités manufacturières ainsi que leur remplacement par une combinaison d'emplois, du secteur des services, exigeant peu de compétences (les "McJobs") et, dans une moindre mesure, d'emplois hautement spécialisés du secteur de la haute technologie. (Les derniers documents publiés au Canada sont moins polémiques et plus prudents dans l'interprétation des données.)

C'est, en partie, à cause de la gamme étendue d'indicateurs statistiques qui ont été utilisés que l'on a observé ou que l'on n'a pas observé le phénomène du recul de la partie du milieu. A leur tour, les documents publiés ont eu tendance à brouiller les concepts de l'inégalité et de la polarisation - ce dernier terme étant celui que nous employons pour parler de la classe moyenne qui recule.

Dans le présent article, on a évalué l'existence du phénomène au Canada dans plusieurs optiques et on a tenté de déterminer les tendances associées qui peuvent avoir joué un rôle déterminant. Contrairement à l'inégalité du revenu entre les familles, qui est demeurée stable au cours des deux dernières décennies, on relève des preuves de polarisation accrue. La polarisation a augmenté tant dans l'optique des familles et de leur revenu total avant impôt que pour les travailleurs individuels et leur revenu du travail.

Dans le cas des familles, l'augmentation de la polarisation semble associée à des changements dans la taille et dans la composition de la famille. Ces derniers changements sont surtout dus au déclin de la fécondité et à l'augmentation du divorce.

Pour les travailleurs individuels, les principaux facteurs semblent être l'augmentation du taux d'activité féminin ainsi que l'augmentation du travail à temps partiel (c.-à-d. pendant moins d'une semaine ou d'une année complète). D'autres facteurs tels que les changements dans la composition de la population active selon l'âge, la profession et la branche d'activité ne semblent pas expliquer l'augmentation de la polarisation.

Bien entendu, ces résultats sont provisoires, particulièrement à cause des détails limités et, dans certains cas, de la qualité des données d'enquête plus anciennes. La publication de résultats plus définitifs doit attendre l'analyse des données du recensement.

BIBLIOGRAPHIE

Atkinson, A.B. (1970), "On the Measurement of Inequality", *Journal of Economic Theory*, Vol. 2.

Beach, C.M. (1988), "The 'Vanishing' Middle Class?: Evidence and Explanations", *Queen's Papers in Industrial Relations*, Industrial Relations Centre, Queen's University at Kingston.

Cowell, F.A. (1977), *Measuring Inequality*, Oxford, Philip Allan Publishers.

Economic Council of Canada (1987), *Innovation and Jobs in Canada*, Catalogue No. EC22-141/1987E, Canadian Government Publising Centre, Supply and Services Canada, Ottawa.

Economic Council of Canada (1990), *Good Jobs, Bad Jobs - Employment in the Service Economy*, Catalogue No. EC22-164/1990E, Canadian Government Publising Centre, Supply and Services Canada, Ottawa.

(l'implosion démographique) après 1966 qui se manifeste par un ralentissement de la croissance de ce groupe d'âges au début des années 1980 et les perspectives d'emploi médiocres pour les jeunes au cours des années 1980, ce qui décourage l'entrée dans la population active. Le groupe des personnes de 25 à 34 ans croît en proportion de l'ensemble de la population active, ce qui reflète l'entrée de la génération de l'explosion démographique, alors que la croissance du groupes des personnes de 35 à 49 ans est fort probablement due à des augmentations du taux d'activité féminin.

En fonction des revenus relatifs dans le graphique 2.b, c'est chez les jeunes que l'on retrouve les plus bas niveaux, ces derniers s'établissant à environ la moitié de la moyenne, alors que l'on retrouve les plus hauts niveaux (environ 125%) chez les personnes du groupe d'âges de 35 à 49 ans. Il n'y a pas de tendances très marquées dans les revenus moyens relatifs parmi les quatre groupes d'âges.

On relève des tendances virtuellement parallèles de l'inégalité du revenu pour les quatre groupes d'âges et les niveaux sont assez semblables dans le graphique 2.c. Ainsi, il est peu probable que les changements dans la structure par âge de la population active réelle expliquent la petite tendance globale à la hausse de l'inégalité.

De même, dans le graphique 2.d, les quatre groupes d'âges montrent tous des tendances généralement parallèles pour la polarisation, bien que les niveaux soient fort différents. C'est chez les jeunes que l'on retrouve la plus faible proportion avec des niveaux de revenu moyens et la plus haute inégalité mesurée pour ce qui est du revenu du travail. C'est la situation inverse qui prévaut pour le groupe des personnes de 25 à 34 ans. La proportion des revenus de niveau moyen qui diminue se retrouve dans les quatre groupes d'âges de sorte, qu'encore une fois, la tendance globale relative à la polarisation observée plus tôt ne peut être expliquée par des changements dans la structure par âge.

Sexe et statut de travailleur à plein temps ou à temps partiel: Nous passons ensuite à une désagrégation différente, de quatre façons, de la population des PVPA, comme les figures 3.a à 3.d le montrent. Cette fois, la population a été divisée selon le sexe et selon que la personne a travaillé ou non pendant des semaines complètes (plus de 35 heures par semaine habituellement) et pendant tout l'année (50 semaines et plus). Si l'une ou l'autre de ces conditions n'était pas satisfaite, le particulier était classé comme un travailleur à temps partiel. L'augmentation du taux d'activité féminin est manifeste dans le graphique 3.a. De plus, tant pour les hommes que pour les femmes, les travailleurs à temps partiel représentent une proportion croissante de la population active.

La proportion d'hommes travaillant à plein temps qui diminue et la proportion croissante de travailleurs à temps partiel étaient associées à une augmentation du revenu moyen relatif des hommes travaillant à plein temps et, en termes relatifs, à une augmentation encore plus grande des gains moyens relatifs des femmes employées à plein temps, comme on le voit dans le graphique 3.b. Alors qu'il y a eu une augmentation tant dans la proportion des travailleuses à temps partiel que dans leurs gains moyens relatifs, bien que le point de départ ait été faible, leurs homologues masculins ont connu une diminution de leurs gains relatifs du début au milieu des années 1980.

Contrairement à la désagrégation en fonction de l'âge utilisée dans l'ensemble précédent de graphiques, pour le présent ensemble de graphiques, les tendances relatives à l'inégalité intragroupe montrées dans le graphique 3.c sont faibles par rapport aux différences entre les groupes, particulièrement dans le cas du travail à temps plein et à temps partiel. Ainsi, les changements dans la composition de la population peuvent expliquer une partie de la tendance globale dans l'inégalité du revenu du travail. Cette tendance est faible mais croissante, alors que l'inégalité intragroupe a tendance à diminuer pendant la première période et qu'elle est mixte pendant la seconde. Ainsi, la tendance globale à la hausse de l'inégalité doit, au moins en partie, être attribuée à un déplacement dans la population active qui compte maintenant une proportion plus élevée de travailleurs à temps partiel dont les gains sont à la fois plus faibles et plus inégalement répartis.

De même, ces changements dans la composition de la population active semblent expliquer une partie de la tendance relative à la polarisation. Dans les groupes d'hommes et de femmes travaillant à temps partiel, il n'y a virtuellement pas de tendance relative à la polarisation; pas plus qu'on ne trouve de tendance très prononcée pour les hommes travaillant à plein temps, comme cela est montré dans le graphique 3.d. La seule tendance évidente se retrouve chez les femmes travaillant à plein temps et alors, seulement au cours des deux dernières périodes. Ainsi, la proportion croissante de la population active réelle qui travaille à temps partiel semblerait expliquer une partie de la tendance globale vers une polarisation accrue des revenus du travail.

Branche d'activité: Les graphiques 4.a à 4.d présentent des résultats correspondants selon de grands groupes de branches d'activité. Ces groupes sont les meilleurs que l'on peut définir, de façon cohérente, pour les quatre enquêtes et il faut faire preuve de prudence à cause du grand nombre de branches d'activité "non codées" dans les données de 1967 et de 1973. Il n'est pas surprenant de constater que la croissance la plus rapide s'est produite dans le secteur des services (commerce de gros et de détail, services personnels et services aux entreprises, intermédiaires financiers, distribution), alors que le secteur où il y a eu le plus grand déclin (à part les branches d'activité "non codées") est celui des activités manufacturières. Cependant, si l'on ne tient pas compte des plus petits groupes, le secteur primaire (agriculture, exploitation forestière, pêche, mines) et la construction, on n'a pas trouvé de tendances significatives dans les gains moyens relatifs, pas plus qu'on n'a relevé de différences importantes dans les niveaux.

Tous les groupes de branches d'activité montrent des tendances à la hausse dans l'inégalité du revenu du travail, milliers de dollars. La ligne supérieure montre le revenu du travail total moyen pour les FR comptant au moins un PVPA. Ces montants sont de 50% à 75% supérieurs aux revenus T,PVPA moyens parce que les gains des conjoints et des enfants ont été groupés. Alors que les courbes T,PVPA et Avl,PVPA montrent une tendance temporelle presque identique, la courbe T,FR-PVPA montre une croissance plus abrupte jusqu'en 1981. Cela est presque certainement dû aux taux d'activité croissants des femmes au cours de la période. Cependant, cette tendance n'a pas suffi à empêcher la stagnation des revenus du travail moyens parmi les FR comptant au moins un PVPA pendant la période allant de 1981 à 1986.

Finalement, la courbe la plus basse du graphique 1.b (Avl,Part.) montre le revenu total moyen pour toutes les personnes de 15 ans et plus, pas seulement pour les PVPA. Elle se trouve au dessous de la courbe Avl,PVPA parce que les particuliers dont la participation au marché du travail n'est pas très importante tendent à avoir des revenus beaucoup plus faibles, en moyenne -- soit qu'ils n'en ont aucun, soit qu'ils reçoivent des montants modestes comme revenu de placements ou comme revenu qui provient surtout de transferts gouvernementaux. Néanmoins, comme dans le cas du graphique 1.a, les tendances temporelles générales sont compatibles.

Le graphique 1.c porte sur l'inégalité telle que mesurée par l'indice de Gini. Les courbes sont tellement rapprochées que seulement celle du haut et celle du bas sont désignées. Les trois courbes sans désignation se rapportent toutes au revenu total avant impôt -- Avl,FR; Avl/UEA,FR et Avl, PVPA. Bien que le choix des unités déclarantes et du concept du revenu aient un effet sur le niveau d'inégalité mesuré, il a un impact négligeable sur la tendance apparente. Pour les courbes Avl, il n'y a pas de tendance significative relative à l'inégalité; elle est généralement constante comme on l'a conclu auparavant dans Wolfson (1986a). Pour les courbes T, on remarque une petite tendance à la hausse, comme on l'a mentionné lors de la discussion portant sur le tableau 2 ci-dessus.

Finalement, le graphique 1.d montre diverses courbes pour les tendances relatives à la polarisation, mesurée comme pourcentage de la population dont le revenu est compris entre 75% et 150% de la médiane. Encore une fois, comme dans les tableaux 1 et 2, on relève une nette tendance à la baisse dans la proportion d'unités à revenu moyen, avec certaines différences dans les niveaux correspondant aux concepts différents.

Il y a, cependant, une exception importante. Au niveau des familles, la tendance disparaît quand le revenu familial total est rajusté pour tenir compte des variations dans la taille des familles (la courbe désignée par Avl/UEA,FR). Par contre, il est clair que le revenu tiré du travail est devenu plus polarisé, que nous considérions les PVPA pris individuellement (la courbe désignée par T,PVPA), ou regroupés en familles de recensement (T,FR-PVPA). Il semble aussi que le revenu total des familles (Avl,FR) soit devenu plus polarisé, comme cela est aussi montré dans le tableau 1 ci-dessus.

L'implication est donc que les déclins dans la taille des familles ont été associés avec des changements dans les revenus des familles d'une façon qui a un effet compensatoire, du point de vue de la polarisation. Le "recul de la classe moyenne" dans l'optique des familles en utilisant le revenu total est apparemment un artefact dû au fait que l'on n'a pas tenu compte des changements systématiques qui se sont produits dans la taille des familles.

Par contre, les tendances relatives à la polarisation au niveau des particuliers demeurent claires et sont associées à des changements dans le marché du travail.

E. EXPLICATIONS POSSIBLES — POLARISATION DU REVENU DU TRAVAIL

Nous passons maintenant à un examen d'autres facteurs qui pourraient expliquer la polarisation accrue du revenu du travail des ouvriers. L'"histoire" que l'on raconte souvent à propos de la classe moyenne qui recule se rapporte à des concepts tels que la "désindustrialisation" et la "déqualification". Malheureusement, les données disponibles à partir des enquêtes sur les finances des consommateurs que l'on utilise pour effectuer la présente analyse ne sont pas très appropriées pour évaluer de tels concepts. Nous serons donc limités à des variables plus conventionnelles -- âge, sexe, travail à plein temps ou à temps partiel, branche d'activité et profession. De plus, pour les deux dernières variables, nous devrons nous contenter de certaines classifications très grossières à cause des limitations qui s'appliquent aux données.

L'approche sera la même pour chaque groupe de variables. La population des personnes qui sont des PVPA sera divisée en 4 à 6 groupes mutuellement exclusifs, puis nous examinerons quatre graphiques. Dans tous les cas, l'année civile figurera le long de l'axe horizontal, exactement comme pour les graphiques 1.a à 1.d ci-dessus. Le premier graphique montre la répartition proportionnelle de la population parmi les divers groupes et comment elle a évolué dans le temps. Le deuxième graphique montre comment le revenu moyen pour chaque groupe se compare au revenu moyen global, le revenu moyen relatif, exprimé en pourcentage. Les deux derniers graphiques montrent les mêmes mesures de l'inégalité et de la polarisation qu'auparavant -- l'indice de Gini et la part de la population donnée dont les revenus se trouvent entre 75% et 150% de leur médiane.

Structure par âge Les graphiques 2.a à 2.d montrent les tendances dans les quatre variables qui viennent d'être décrites pour les quatre groupes d'âges: 15-24, 25-34, 35-49 et 50 ans et plus. Le plus vieux groupe d'âges diminue au cours des deux décennies comme proportion de la population active réelle, ce qui reflète la diminution du taux d'activité des hommes plus âgés. La proportion des jeunes, par contre, augmente légèrement de 1967 à 1973, mais elle diminue par la suite. Cela reflète probablement la chute brusque de la fécondité du revenu médian.

Ces tendances relatives à la polarisation sont presque certainement statistiquement significatives tant pour la combinaison particuliers/revenu du travail que pour celle des familles/revenu total. On peut démontrer, en pratique, que les tendances relatives à la polarisation ne correspondent pas nécessairement à des tendances dans le même sens pour ce qui est de l'inégalité du revenu; ce sont, en fait, des concepts distincts.

Alors que les statistiques présentées dans les tableaux 1 et 2 ont ramené les concepts du revenu moyen, de l'inégalité et de la polarisation à seulement 14 nombres, il y en a encore trop pour effectuer des analyses plus détaillées. Ainsi, à des fins de présentation, nous concentrerons notre attention sur l'indice de Gini utilisé comme mesure de base de l'inégalité et sur la part de la population dont les revenus sont compris entre 75% et 150% de la médiane comme notre indice de base de la polarisation. Nous avons quand même examiné la gamme plus étendue de statistiques dans tous les cas dont nous traiterons afin de nous assurer que la seule statistique sommaire présentée pour chaque concept représentait fidèlement les tendances de base.

D. EXPLICATIONS POSSIBLES - UNITÉS DÉCLARANTES ET CONCEPT DU REVENU

Il se peut que les tendances relevées dans les tableaux 1 et 2 ci-dessus soient une sorte d'artefacts statistiques résultant de nos choix particuliers d'unités déclarantes du revenu et de définitions du revenu. Avec une exception importante, nous démontrerons dans la présente section que ce n'est pas le cas.

Jusqu'ici, nous avons concentré notre attention sur deux unités déclarantes du revenu -- les familles de recensement (FR) qui comprennent l'époux et (ou) l'épouse et les enfants jamais mariés vivant dans le même logement et les participants véritables à la population active (PVPA). Une autre définition plus générale et très utilisée de la famille est celle des familles économiques (FE), définie comme toutes les personnes apparentées vivant dans le même logement. Nous pouvons aussi définir le sous-ensemble des FR qui comprennent au moins un PVPA, familles que nous désignerons à l'aide du sigle FR-PVPA. Finalement, nous utiliserons l'abréviation Part. pour désigner la population de tous les particuliers de 15 ans et plus, qu'ils soient ou non des PVPA.

Jusqu'ici, nous ne nous sommes attachés qu'à deux concepts du revenu --le revenu total ou avant impôt et le revenu du travail. Nous représenterons ces revenus par AvI et T respectivement. Le revenu après impôt (ApI) est une autre définition du revenu qui présente un intérêt général. Nous aurions fait un plus grand usage de ce concept sauf que, malheureusement, les données de 1967 ne renferment pas d'estimations de l'impôt sur le revenu payé.

Finalement, dans le cas des familles, nous avons généralement employé le revenu par famille. Cependant, cela est presque certainement une mauvaise façon de comparer les situations financières de familles de taille différente. Une façon de tenir compte des différences dans la taille des familles consiste à utiliser le revenu divisé par une échelle d'équivalence. Il s'agit d'une échelle de facteurs numériques que l'on peut grossièrement interpréter comme les besoins en revenu relatifs des familles de taille différente.

Le choix des échelles d'équivalence soulève une vive controverse, voir, par exemple, la discussion à ce sujet dans Wolfson et Evans (1989). Nous avons choisi d'utiliser une échelle qui attribue un poids de 1.0 au premier adulte, de 0.4 à chacun des autres adultes dans la famille et de 0.3 aux enfants (sauf dans le cas du premier enfant dans une famille monoparentale à qui l'on attribue un poids de 0.4). Ces poids représentent une échelle d'unités équivalentes à des adultes (UEA). Ainsi, un couple marié avec deux enfants a un poids de 2.0.

Avec de telles échelles d'UEA, nous pouvons analyser les répartitions du revenu des familles où l'on divise tout d'abord le revenu par le nombre d'UEA dans la famille. Ainsi, un couple marié avec deux enfants et un revenu total de $25,000 serait traité comme une famille avec un revenu de $12,500 par UEA.

On pourrait s'attendre à ce que ce genre de rajustement à l'aide des UEA ait un effet important à cause des tendances significatives au cours des deux dernières décennies en ce qui a trait à la diminution du taux de fécondité et à l'augmentation du taux de divorce et donc d'une diminution de la taille moyenne des familles.

Les graphiques 1.a à 1.d montrent la sensibilité des résultats de base présentés dans les tableaux 1 et 2 aux divers choix d'unités déclarantes du revenu et de concepts du revenu qui viennent d'être définis. Le graphique 1.a montre les tendances dans le revenu réel moyen des familles pour quatre cas différents. Les revenus totaux moyens les plus élevés se retrouvent pour les FE, ce qui n'est pas surprenant parce que ces familles comprennent plus de membres et donc plus de bénéficiaires d'un revenu. Parmi les FR, le revenu après impôt est, en moyenne, de $5,000 inférieur au revenu avant impôt. Finalement, le revenu total moyen par UEA pour les FR était environ 60% du revenu total moyen par FR. Cependant, peu importe le concept du revenu ou la définition de l'unité déclarante, le comportement historique général est le même que celui que l'on a fait remarquer pour les tableaux 1 et 2 -- une croissance importante du revenu réel à la fin des années 1960 et au début des années 1970, mais une stagnation et un déclin du début au milieu des années 1980.

Le graphique 1.b montre les tendances dans les revenus moyens réels des particuliers. Les données qui se trouvent dans le tableau 2, on s'en souvient, s'appliquaient aux PVPA et à leur revenu du travail (T). La courbe AvI,PVPA montre le revenu total moyen avant impôt pour ces mêmes particuliers; il est plus élevé de quelques

Les statistiques que nous avons choisies d'utiliser pour montrer les tendances relatives à l'inégalité et à la polarisation sont mathématiquement simples afin que les résultats soient aussi clairs et intuitifs que possible. Cependant, le raisonnement à la base du choix particulier d'indicateurs est un peu plus technique, il est donc présenté dans l'annexe.

C. RÉSULTATS GLOBAUX

Le tableau 1 présente, dans l'optique d'une famille, les tendances de base relatives à l'inégalité et à la polarisation pendant près de deux décennies au Canada. La présentation du tableau 2 est identique, mais le tableau se concentre sur les particuliers avec participation non négligeable au marché du travail au cours de chaque année. Plus précisément, le tableau 1 examine les familles de recensement et leur revenu total. Le tableau 2, par contre, ne porte que sur les particuliers âgés de 15 ans et plus qui ont reçu, au cours de l'année, un revenu du travail dont le montant est supérieur à 2.5% du salaire moyen pour l'année. Nous appelons ces personnes des "participants véritables à la population active" ou PVPA. En gros, ils doivent avoir travaillé au moins une semaine, à plein temps, à un taux de rémunération égal au salaire moyen, ou au moins deux semaines, à plein temps, au salaire minimum.

Ces deux tableaux représentent donc deux grandes optiques de la répartition des revenus -- la première se concentrant sur les familles et sur leur revenu provenant de toutes les sources, pas seulement du travail et la seconde se concentrant sur les PVPA et sur leur revenu du travail (y compris le travail autonome).

Dans tous les cas, les données sont tirées d'analyses spéciales des résultats des enquêtes sur les finances des consommateurs. Les années particulières examinées ont été choisies afin de fournir la série chronologique la plus longue possible pour laquelle les données sont disponibles et cohérentes, les tailles des échantillons sont grandes et l'économie était à des points à peu près semblables du cycle économique. Ce sont exactement les mêmes données sous-jacentes qui ont été utilisées par Picot et coll. (1990). La présente analyse et celle de Picot et coll. sont complémentaires, parce que cette dernière porte sur un sous-ensemble de PVPA qui ont travaillé à plein temps pendant toute l'année. Les données sur les PVPA présentées ici sont à la base de celles qui ont été utilisées par le Conseil économique du Canada (1989).

Chaque tableau comprend trois groupes de statistiques. On trouve sur les deux premières lignes la moyenne et le point milieu de la répartition des revenus pour chaque année - les revenus moyen et médian exprimés tous deux en dollars constants de 1986. Bien que les chiffres aient été arrondis au $50 le plus rapproché, la variabilité d'échantillonnage est telle qu'ils ne sont réellement précis qu'à environ $500 près. Tant dans l'optique des familles et du revenu total que dans celle des particuliers et du revenu du travail, les revenus ont augmenté le plus rapidement à la fin des années 1960 et au début des années 1970. La croissance des revenus moyens a alors ralenti, puis elle a connu une période de stagnation et même de déclin du début au milieu des années 1980.

Le second groupe de chiffres se rapporte à l'inégalité du revenu. Ce sont les parts de revenu qui reviennent à chaque quintile le long du spectre des revenus et l'indice de Gini, un indice global de l'inégalité. Dans l'optique des familles et du revenu total, toutes ces statistiques varient un peu d'une année à l'autre. Cependant, des estimations grossières de la variabilité d'échantillonnage de ces chiffres laissent supposer qu'il n'y a pas de tendances statistiquement significatives. Par exemple, alors que la part du revenu total qui revient au cinquième supérieur des familles varie jusqu'à 1.5%, l'intervalle de confiance à 95% est probablement d'au moins deux points. Love et Wolfson (1976, annexe 2) dans un contexte semblable estiment que l'erreur-type relative de l'indice de Gini varie entre 1.5% et 3.5% (c.-à-d. des valeurs de 0.6% à 1.4%), selon la taille de l'échantillon. Ainsi, les différences d'au plus 1.5% (41.3 - 39.8) dans l'indice de Gini ne sont vraisemblablement pas statistiquement significatives.

La question d'une tendance dans l'inégalité du revenu du travail des particuliers constitue plus un cas limite. Les chiffres montrent effectivement une augmentation de l'inégalité qui pourrait bien être statistiquement significative.

Par opposition à l'absence de tendance dans l'inégalité, les données montrent effectivement une tendance évidente vers une polarisation accrue - mesurée par un déclin dans le nombre de familles et de particuliers avec des revenus "près du niveau moyen". Par exemple, le tableau 1 montre une diminution d'environ un septième du nombre de familles à revenu moyen qui passe de 37.2% à 31.7% pour les familles dont le revenu est compris entre les trois quarts du revenu familial médian et une fois et demie ce revenu.

Le tableau 2 montre, sur la dernière ligne, un déclin, encore plus marqué, d'un cinquième dans le nombre de travailleurs avec un revenu du travail de niveau moyen qui est passé de 39.3% à 30.8% pour les personnes dont le revenu est compris entre les trois quarts du revenu du travail des particuliers médian et une fois et demie ce revenu. Si l'on étudie un peu plus les données, on constate que ce déclin s'est produit dans des proportions à peu près égales tant dans le groupe dont les revenus sont compris entre 75% et 125% de la médiane que dans celui dont les revenus vont de 125% à 150% de la médiane. De 1967 à 1986, 5.4 des 8.5 points de diminution, du nombre de travailleurs dont les revenus sont compris dans la gamme combinée de 75% à 150% du revenu médian "se sont déplacés" (pas littéralement, parce que les données ne sont pas longitudinales) vers la gamme "plus de 150%" du revenu median, alors que les autres se sont retrouvés dans la gamme des revenus inférieurs à la moitié

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

INÉGALITÉ ET POLARISATION: LA CLASSE MOYENNE RECULE-T-ELLE AU CANADA?

M.C. Wolfson[1]

A. INTRODUCTION

On retrouve depuis longtemps un intérêt général relatif aux tendances de l'inégalité du revenu dans les sociétés comme celle du Canada. L'opinion populaire est bien représentée par l'expression "les riches s'enrichissent et les pauvres s'appauvrissent". Au cours des années 1980, cet intérêt a pris une nouvelle tournure alors qu'aux Etats-Unis on a commencé à parler de la "désindustrialisation". La disparition graduelle d'emplois industriels relativement bien payés remplis par des cols bleus constitue une partie de l'hypothèse de la désindustrialisation. On pense que ces emplois sont remplacés par une combinaison d'emplois, peu payés et exigeant peu de compétences, du secteur des services (les "McJobs") et un plus petit nombre d'emplois de cols blancs bien rémunérés et qui exigent beaucoup de compétences, comme les emplois d'analystes en informatique.

On a vu ce dernier phénomène comme étant à l'origine d'un "recul de la classe moyenne" qui commence. Une avalanche d'articles ont été publiés sur ce sujet, Kuttner (1983) en étant un des premiers exemples. Plus récemment, ces documents ont été examinés dans Loveman et Tilley (1988), où les contributions canadiennes étaient les suivantes: Conseil économique (1987), Myles (1987), Leckie (1988) et Picot et coll. (1990). Un bon nombre des analyses dans ce domaine ont donné lieu à beaucoup de controverses et attiré beaucoup d'attention surtout à cause des choix différents de concepts et de définitions.

Le présent article a pour objectif de présenter les faits de base relatifs aux tendances, en matière d'inégalité et de polarisation, au Canada, au cours des deux dernières décennies. La polarisation est le concept que nous utilisons pour capter la notion de la classe moyenne qui recule, il s'agit d'un concept différent de celui de l'inégalité comme Love et Wolfson (1976) l'ont fait remarquer.

Avant de considérer les résultats empiriques, nous commencerons par présenter un certain nombre de définitions et de concepts. Pour anticiper les conclusions, nous ne trouvons aucune tendance significative dans l'inégalité du revenu, la même "stase dans le changement" qui a été signalée dans Wolfson (1986a). Cependant, nous trouvons des preuves de polarisation accrue de la répartition des revenus, mais pas pour les raisons auxquelles nous pensions au début.

B. INÉGALITÉ ET POLARISATION DE QUOI POUR QUI

La majorité des analyses des tendances dans l'inégalité du revenu tendent à tenir compte du revenu total ou après impôt des familles. Par contre, un bon nombre des analyses portant sur la "classe moyenne qui recule" examinent les gains au titre des salaires des travailleurs et, parfois, les titres de professions, par exemple, les taux de croissance relatifs des professions avec le plus haut et le plus bas salaire (p. ex., Rosenthal, 1985). Ainsi, les discussions portant sur l'inégalité et sur la polarisation peuvent devenir embrouillées si l'on utilise des unités d'analyse et des mesures de la position économique qui varient d'une étude à l'autre.

Dans la présente analyse, on tiendra compte du revenu total et du revenu disponible ainsi que du revenu provenant des salaires et des traitements seulement (le revenu "du travail"). On considérera aussi les familles en général et les particuliers avec une participation non négligeable au marché du travail. De plus, on examinera des rajustements au revenu des familles afin de tenir compte du changement dans la taille moyenne et dans la composition des familles. Ainsi, nous utiliserons diverses optiques, de façon cohérente, afin d'analyser les tendances relatives à l'inégalité et à la polarisation.

La variété des mesures statistiques utilisées dans les différentes études constitue une autre source de confusion relative aux analyses de l'inégalité et de la polarisation. Le texte d'une étude peut parfois porter sur l'inégalité alors que les statistiques dans les tableaux qui accompagnent le texte sont mathématiquement incompatibles avec ce concept. Dans d'autres cas, le texte porte sur la classe moyenne qui recule alors que les tableaux renferment des mesures de l'inégalité. Ces confusions ont une importance fondamentale.

En ce qui nous concerne, nous utiliserons deux grands groupes de statistiques tirées des répartitions des revenus à analyser -- celles qui se rapportent à l'inégalité et celles qui se rapportent à la polarisation. Bien que cela ne soit pas encore généralement admis, il s'agit de concepts distincts. Il se peut qu'une répartition des revenus soit plus égale qu'une autre, tout en montrant une plus grande polarisation. Intuitivement, l'inégalité porte sur la gamme de différences parmi toute la population, alors que le concept de polarisation reflète dans quelle mesure des particuliers ou des familles tendent à se rassembler en deux groupes distincts le long du spectre des revenus.

[1] M.C. Wolfson, Division des études sociales et économiques, Statistique Canada, Ottawa (Ontario), K1A 0T6.

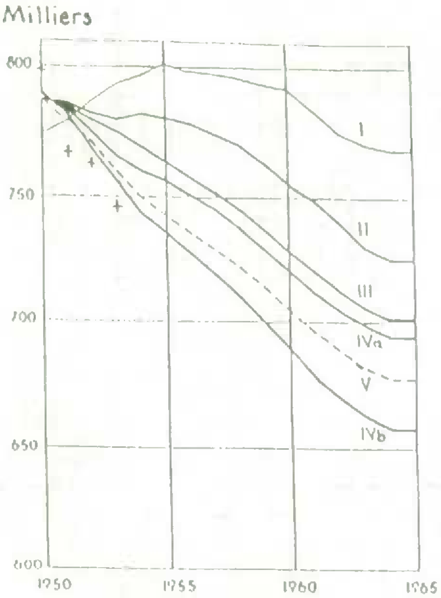


Figure 11 - France. Prévisions de naissances légitimes selon diverses méthodes.
Source : L. Henry. Perspectives de naissances après une perturbation de la natalité. Communication au Congrès de l'U.I.E.S.P. à Rome (1955).

Séries chronologiques concomitantes.

La mise en rapport d'une série chronologique relative à des événements démographiques avec d'autres séries relevant de la vie économique, politique et sociale, est un point clef de la recherche causale en démographie. Cet aspect de la recherche ne mobilise aucune technique spécifique à la démographie et nous n'envisagerons pas ici cette question. Les résultats des travaux effectués dans ce sens restent minces comme en témoigne le bilan qu'en a dressé H. Leridon (Cf. bibliographie), au sujet de la fécondité.

Vues générales.

Au terme de ces considérations, on peut avoir le sentiment d'un discours par touches, l'accent étant mis en priorité sur l'utilisation des séries chronologiques pour la prévision. Comme il a été dit en introduction, parler des séries chronologiques c'est parler de la substance première de l'analyse démographique et c'est reconnaître qu'au total, en raison des habitudes relativement récentes de collecte suffisamment riche, il existe une certaine indigence de l'information dans ce domaine. Et cette indigence a sûrement les conséquences les plus fâcheuses en matière de prévision et c'est pourquoi nous avons mis l'accent sur cet aspect de l'utilisation des dites séries.

BIBLIOGRAPHIE

Certains développements qui précèdent s'appuient sur des résultats ayant fait l'objet de diverses publications.

R. Pressat - *Distorsions introduites par une vision transversale des phénomènes en démographie*. Communication au Congrès de l'Institut international de statistique, Vienne, 1973.

G. Calot et R. Nadot - *Combien y aura-t-il de naissances dans l'année ?* Population, numéro spécial, 1977.

J.P. Sardon - *L'analyse démographique conjoncturelle : réflexions méthodologiques*. Communication au XXème Congrès général de l'U.I.E.S.P., non publiée (Florence, 1985).

J.P. Sardon - *The short-term forecast of rates : Longitudinal or transversal perspective* ? Materialen zur Bevölkerungswissenschaft, Heft 49, BIB, Wiesbaden, 1986.

L. Henry et H. Gutierrez - *Qualité des prévisions démographiques à court terme*. Population, n° 3, 1977.

L. Henry - *Perspectives de naissances après une perturbation de la natalité*. Communication au Congrès général de l'U.I.E.S.P., Rome, 1954.

R. Pressat - *Un essai de perspectives de ménages*. Congrès international de la population, Vienne, 1959.

H. Leridon - *Natalité, saisons et conjoncture économique*. Cahier de Travaux et documents de l'INED, n° 66, Paris, 1973.

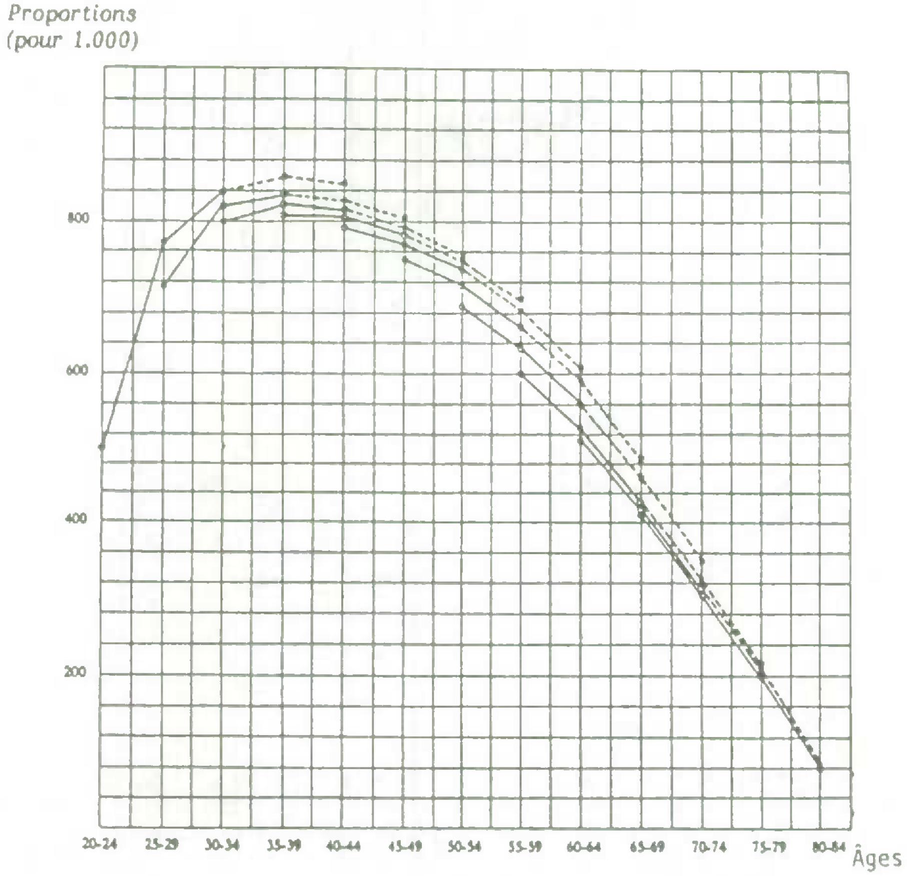


Figure 10 - France. Proportions de femmes mariées en 1947, 1952, 1957 et perspectives pour 1962, 1967.
Source : R. Pressat. Un essai de perspectives de ménages. Communication au Congrès général de l'U.I.E.S.P. à Vienne (1959).

La question ne se pose guère en termes d'analyse de la situation passée, le démographe sachant donner tout son sens à chaque série spécifique. Il en va différemment lorsqu'il doit faire un choix lors de la mise en oeuvre des prévisions. Nous rapportons pour illustrer notre propos un exemple quasiment historique qui a fait l'objet de nos tout premiers travaux sous la direction de L. Henry : la prévision des naissances légitimes dans une hypothèse de stabilité des indices dans le futur, à savoir :

I. - taux de fécondité générale par âge (avec adoption d'une proportion fixe de naissances illégitimes) ;
II. - taux de fécondité par durée de mariage et âge au mariage ;
III. - taux de fécondité par durée de mariage seule ;
IV. - taux de fécondité par nombre d'enfants déjà nés du mariage actuel et intervalle écoulé soit depuis le mariage, soit depuis la naissance précédente. Deux séries IV_a et IV_b découleront de cette approche compte tenu des calendriers des naissances différents qui ont été adoptés.
V. - taux de fécondité des ménages récents (on se trouve placé, à l'époque de ces calculs, en 1950), observés à 0-11 ans de durée et extrapolés au-delà.

La figure 11 montre l'extrême diversité des résultats. Sans doute étions-nous à l'époque au sortir de périodes perturbées avec les séquelles que cela comporte et donc une grande instabilité des indices. Mais il en est toujours un peu ainsi : chaque indice du moment se ressent d'une histoire passée et cela différemment selon sa structure propre, en sorte qu'il se cache dans ce que l'on projette, sous couvert d'un langage analogue, des hypothèses quelque peu dissemblables.

Ce processus peut être remis en oeuvre avec les proportions que représentent les autres états matrimoniaux (veufs et veuves, divorcé(e)s, et maintenant concubins) ; il restera à assurer la cohérence des résultats ainsi établis individuellement pour que la totalité de ces états à un âge donné coïncide avec l'effectif de la population à cet âge (une réduction proportionnelle y pourvoira).

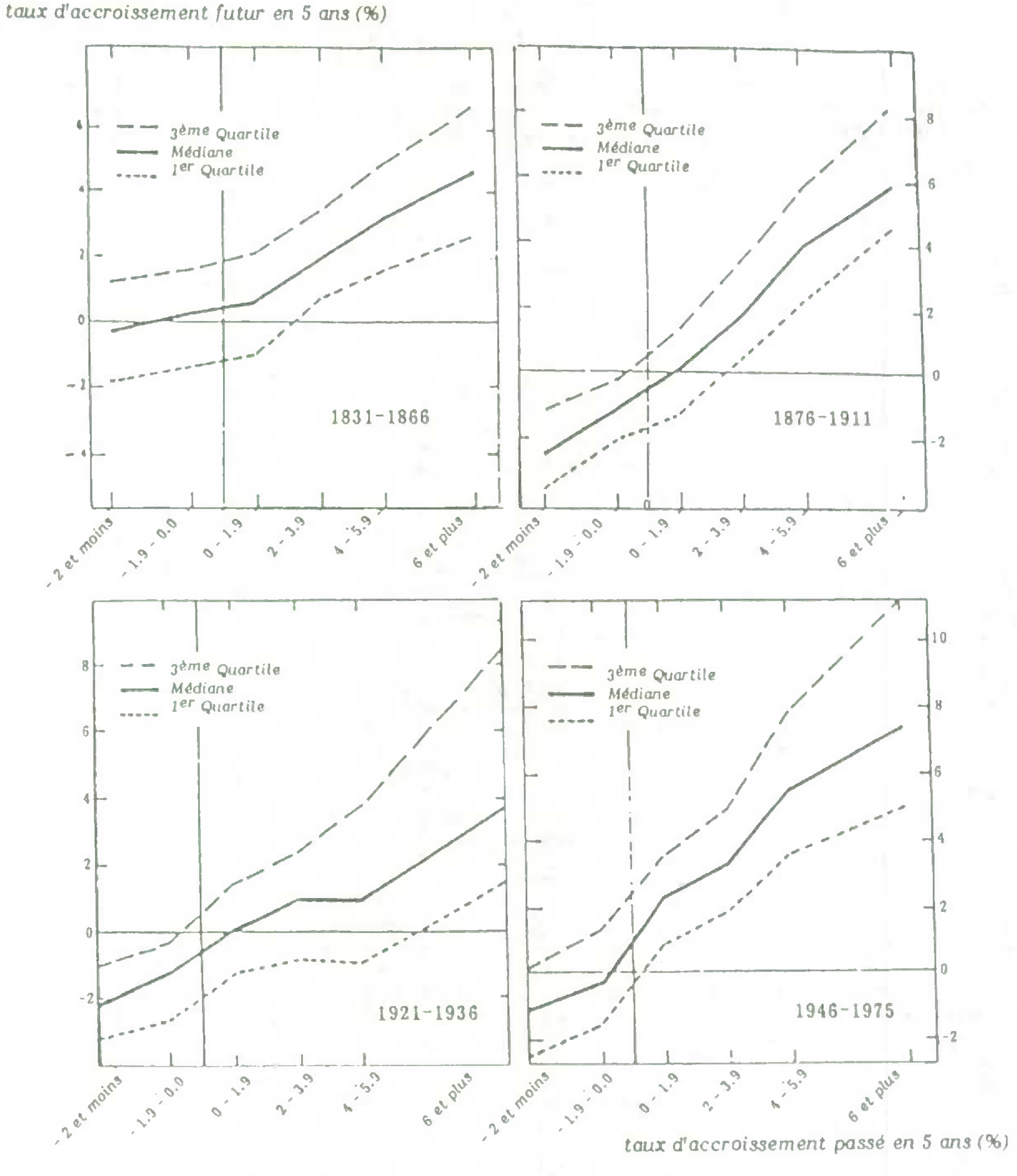


Figure - 9 Départements français. Médianes et quartiles du taux d'accroissement futur en 5 ans (taux observé) suivant le taux d'accroissement passé en 5 ans (taux prévu).

Source : L. Henry et H Gutierrez. Qualité des prévisions démographiques à court terme. Population, n° 3, 1977.

De la diversité des séries chronologiques.

En face de problèmes d'analyse comme de problèmes de prévision, la démographie est souvent aux prises avec plusieurs séries chronologiques relativement au même phénomène : ces diverses séries sont-elles équivalentes en tant qu'instruments de la prévision ?

Compte tenu des meilleurs résultats obtenus lorsque l'on fait intervenir la suite des valeurs à chaque âge, observées au fil du temps, on peut penser que l'on pourra parvenir à de meilleures prévisions en recourant à des ajustements qui tiennent à la fois compte et du passé des générations et de la tendance à âge constant. Cela revient à compléter la formule précédente en introduisant un terme linéaire en t :

$$f(x,t) = a(x) \cdot \sum_{\xi=15}^{t-1} f[\xi, t-(x-\xi)] + b(x)t + c(x). \tag{3}$$

Mais on peut encore reprendre les deux dernières formules en ne prenant en compte que le passé récent des générations. c'est-à-dire en ne sommant la fécondité qu'à partir d'un âge peu antérieur à la date t , notant p cette antériorité, on est conduit aux ajustements s'appuyant sur les formules :

$$f(x,t) = a(x) \cdot \sum_{\xi=t-p}^{t-1} f[\xi, t-(x-\xi)] + b(x) \tag{4}$$

ou

$$f(x,t) = a(x) \cdot \sum_{\xi=t-p}^{t-1} f[\xi, t-(x-\xi)] + b(x)t + c(x).$$

Ces deux dernières formules ouvrent la voie à un nombre considérable d'essais ; faute de pouvoir les examiner dans leur totalité, on peut privilégier ceux répondant à p = 1, autrement dit, les cas où les acquis de l'année précédente, dans les générations en cause, sont pris en compte seuls ou en conjonction avec les résultats au même âge l'année précédente, donc également dans la génération précédente (Cf. J.P. Sardon, 1986).

Deux types de prévisions à moyen terme.

Dans le même esprit que précédemment, mais en s'attachant à un type de prévisions très différent, nous montrerons combien nos faiblesses en matière de choix d'hypothèses pour les prévisions peuvent conduire à des options dont les apparences simplistes n'empêchent qu'elles aient certaines vertus opérationnelles. Il s'agit en la circonstance de prévisions de population pour les départements français.

Face à ces difficultés et dès lors que l'on ne renonce pas à prévoir, une méthode simple s'offre au prévisionniste : reconduire d'une période à la période suivante, objet de la prévision, le taux d'accroissement précédemment observé. Appliquée de 1831 à 1875 -à quelques lacunes près tenant à des anomalies diverses de situations- cette méthode apparemment brutale a conduit aux résultats de la figure 9. Très précisément sur cette figure nous ne mesurons pas l'écart entre prévisions et observations ; au lieu de cela, nous représentons graphiquement les relations qui existent entre les taux d'accroissement de chaque période quinquennale et les taux correspondants de chaque période quinquennale immédiatement postérieure (on note ainsi la correspondance qui existe entre le taux d'accroissement de la population d'un département donné entre 1831-1836 et le taux entre 1836-1841). Si prévoir à 5 ans la population d'un département en reconduisant le taux d'accroissement durant la période de 5 ans qui se termine au point de départ de la prévision était parfaitement correct, les graphiques de la figure 9 devraient se réduire à une ligne diagonale (croissance égale en abscisse -taux passé- et en ordonnée -taux futur).

Il en va différemment. Le tracé en trait plein repose sur les valeurs médianes pour chacune des périodes considérées ; il se situe au-dessous de la première bissectrice pour les trois premières périodes et le plus souvent au-dessus pour la quatrième, signe d'une croissance ralentie dans le premier cas et accélérée dans le second. L'emplacement des tracés correspondant aux premier et troisième quartile donne une idée de l'imprécision découlant de la reconduction du taux passé en tant que moyen de prédire ; les écarts, en valeurs absolues, se situent le plus souvent entre 2 et 4 %.

Aussi insatisfaisant que soit ce résultat, la reconduction du taux d'accroissement de la quinquennie passée représente un progrès par rapport au simple maintien du chiffre de population du précédent recensement ; et l'on vérifiera sur les graphiques que, le plus souvent, on se tromperait plus de 3 fois sur 4 en opérant ainsi.

Ces résultats négatifs prêchent en faveur de la constitution de séries chronologiques aussi étendues et aussi riches de substance que possible. C'est à ce prix que les essais de méthode et la confrontation avec la réalité des résultats qui en découlent pourront être suffisamment nombreux pour dégager les meilleures conduites à tenir en matière de prévision, ce que les seuls analyses et raisonnements a priori ne seraient pas à même de faire.

Nous allons illustrer maintenant un autre type de prévision à moyen terme qui mettra bien en valeur l'intérêt des séries chronologiques. Il s'agit de prévisions de population par état matrimonial. Nous sommes encore dans une situation où l'analyse des processus démographiques à la base de la constitution et de la transformation de ces divers états revêt une telle complexité qu'elle ne saurait conduire à des conduites opérationnelles en matière de prévision. Ici c'est un autre type de séries chronologiques qui va être mis en oeuvre et, dans un premier temps, la série des proportions de personnes mariées selon le sexe et l'âge. L'examen de la figure 10 nous permettra d'être court dans nos explications. Sur le graphique se trouvent repérées les proportions de femmes mariées selon l'âge en 1947, 1952 et 1957 ; les lignes en traits pleins qui joignent les points représentatifs de ces proportions, induisent par la régularité de leur tracé des extrapolations (en tireté) pour les années 1962 et 1967. C'est le recouvrement des tracés relatifs aux observations qui, ici, nous a servi de guide.

on peut ainsi concevoir en tant qu'autre manière d'effectuer une extrapolation, opérer un ajustement sur la base de la relation :

$$f(x,t) = a(x) \,.\, \sum_{\xi=15}^{t-1} f[\xi, t-(x-\xi)] = b(x) \qquad (2)$$

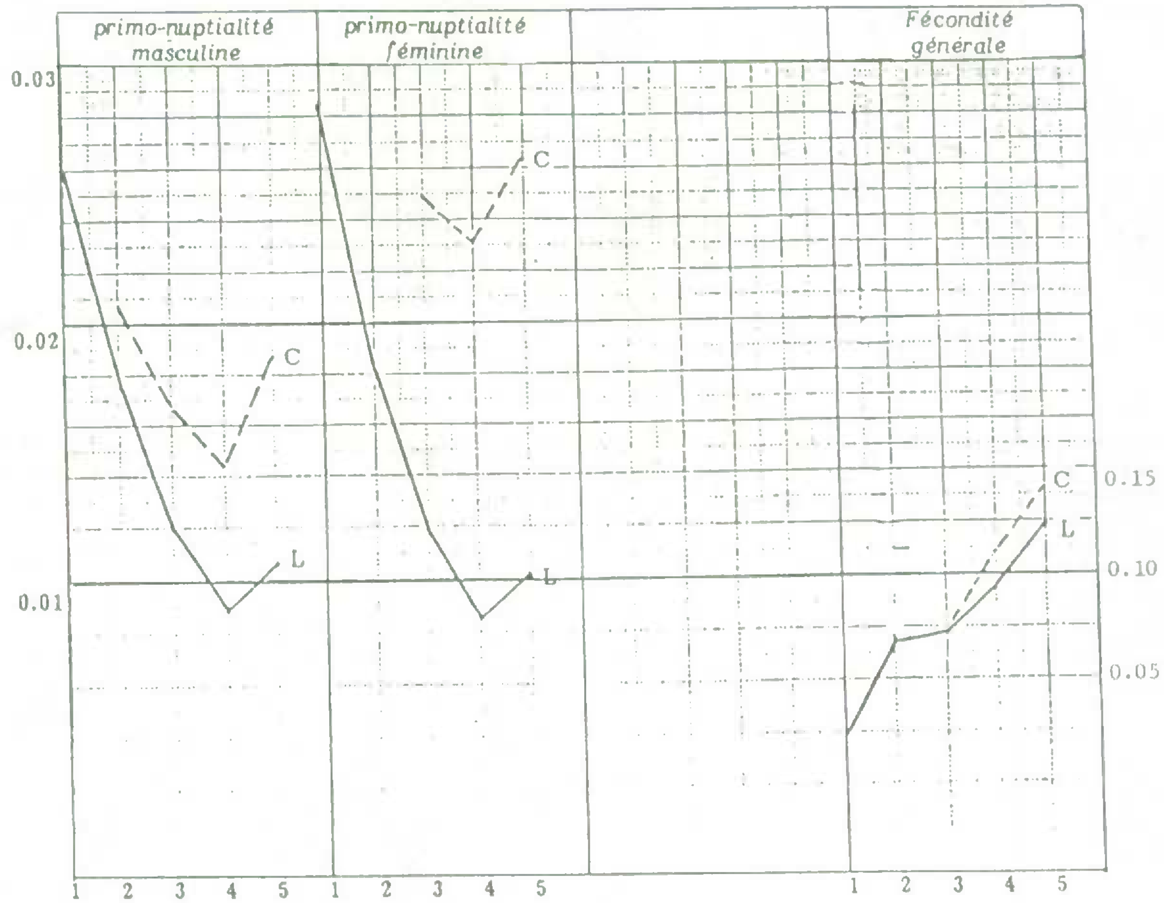
Comme nous l'avons souligné précédemment, l'ajustement peut s'effectuer à partir d'un nombre variable de points ; et, comme précédemment, on a, a priori, le sentiment qu'il existe un nombre optimal suffisamment faible pour tenir compte des tendances récentes mais pas trop pour ne pas induire un ajustement qui se ressentirait beaucoup des variations aléatoires attachées à un nombre trop restreint d'observations.

L'étude dont nous extrayons les développements précédents a été conduite avec les données françaises sur la fécondité et sur la primo-nuptialité masculine et féminine pour la période 1946-1982 ; les prévisions rétrospectives concernent la période 1977-1982, l'extrapolation portant sur une seule année (avec des données arrêtées à 1976 on estime 1977, avec celles arrêtées à 1977 on estime 1978, etc.).

Avec la figure 8, nous pouvons juger de la valeur des deux méthodes présentées. Les extrapolations conduites avec les séries chronologiques, sans prise en compte du passé des générations, sont notées L, celles où ce passé intervient sont notées C.

Curieusement, c'est en faisant abstraction du passé des générations, que ce soit en matière de nuptialité ou de fécondité, que l'on arrive aux meilleurs résultats. De plus, pour ce qui est de la primo-nuptialité, et comme nous l'avons constaté, quand nous avons posé le problème de la projection du nombre des naissances, c'est en prenant en compte un nombre suffisant, mais pas trop important de points du passé (en l'occurrence ici 4) que l'on parvient aux meilleurs résultats. Pour la fécondité, et de façon quelque peu troublante, c'est la simple reconduction de la valeur de l'année précédente qui conduit à la prévision la plus satisfaisante.

Erreurs sur les indices synthetiques (en nombre par personne)



nombres de points du passé pris en compte

Figure 8 -France. Prévisions pour chacune des années 1977 à 1982 (moyennes des résultats).
Source : J.-P. Sardon. L'analyse démographique conjoncturelle : réflexions méthodologiques. Communication au XXème Congrès général de l'U.I.E.S.P., non publiée (Florence, 1985).

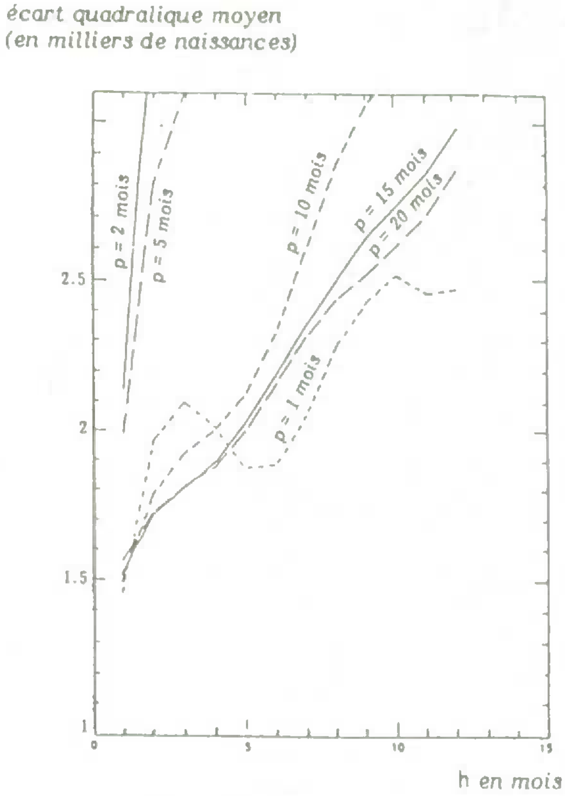


Figure 5 - Données de la figure 3
présentées à p constant.

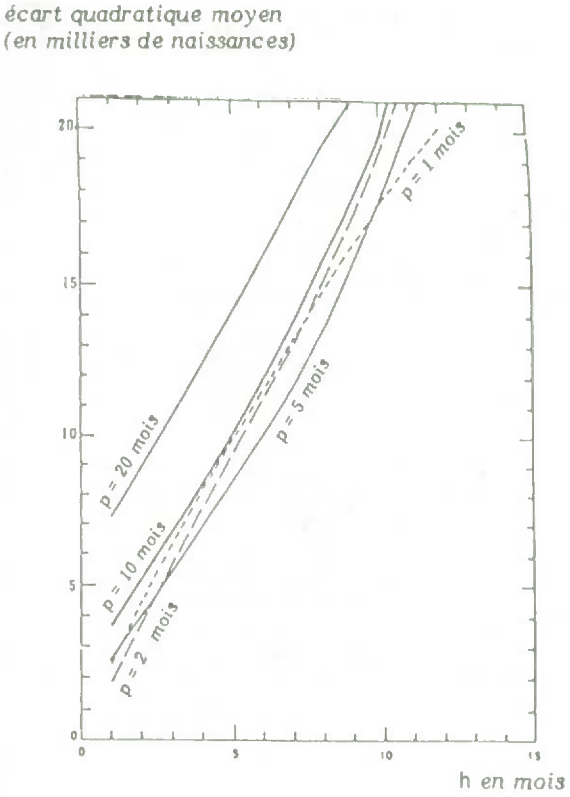


Figure 6 - Données de la figure 4
présentées à p constant.

Prévoir l'indice synthétique de fécondité.

Le recours à l'indice synthétique de fécondité permet une meilleure approche des comportements que l'utilisation des seuls nombres de naissances. Mieux encore, l'examen des taux de fécondité générale par âge entrant dans le calcul de cet indice conduit à une analyse plus approfondie. C'est précisément le problème de l'utilisation des séries chronologiques dans la projection des taux de fécondité que nous allons examiner.

En se référant à la figure 7, le taux à l'âge (atteint) x, à la date t, soit f(x,t), se rapportera à la zone hachurée. Une première approche consistera en un ajustement linéaire :

$$f(x,t)=a(x)t+b(x) \qquad (1)$$

Ce faisant, on ne prend pas en compte le passé des cohortes (ici des générations) en tant que facteur de nature à influer sur le comportement à un âge donné. Faire intervenir la fécondité passée c'est faire intervenir la descendance atteinte au début de l'année t. Cette descendance s'écrit :

$$\sum_{\xi=15}^{t-1} f(\xi,t-(x-\xi))$$

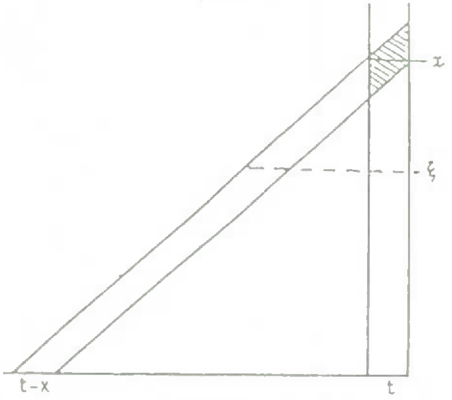


Figure 7

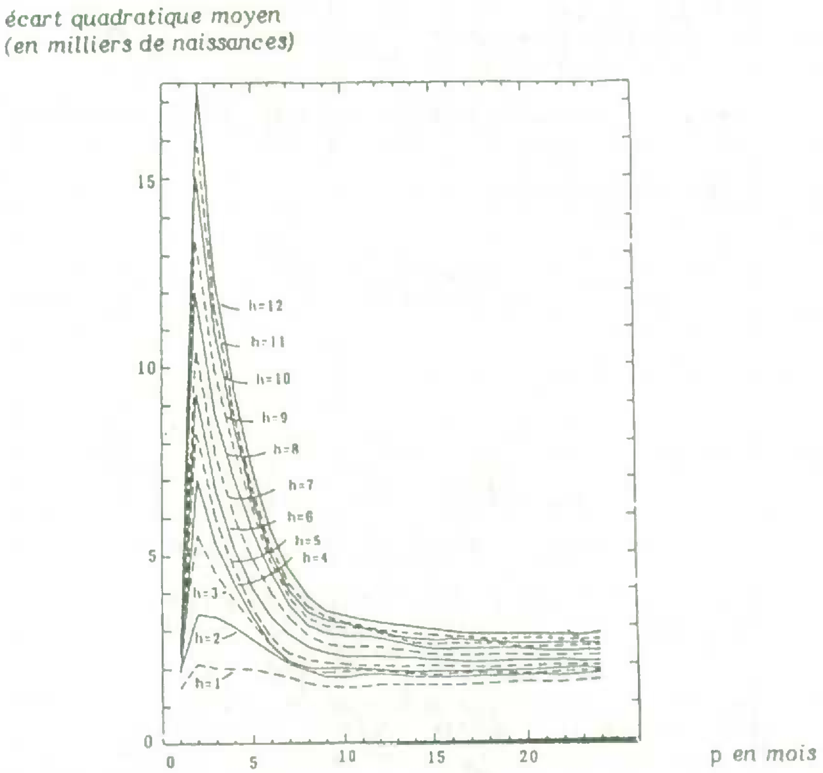


Figure 3 - Ecart quadratique moyen de la prévision des naissances mensuelles désaisonnalisées selon le nombre p de mois de la droite d'ajustement et la distance h du mois sur lequel porte la prévision.
Source : G. Calot et R. Nadot. Combien y aura-t-il de naissances dans l'année ? Population, numéro spécial, 1977.

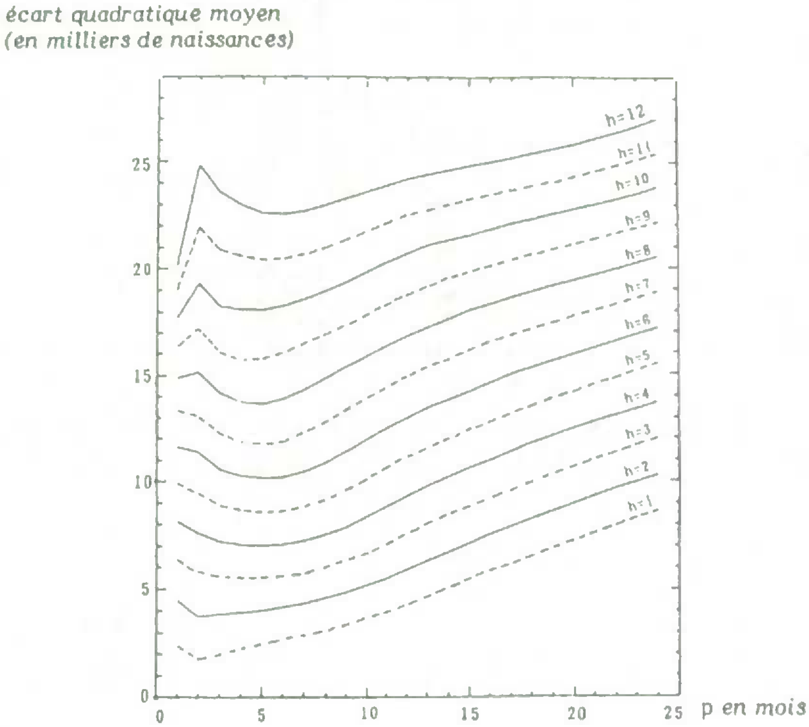


Figure 4 - Ecart quadratique moyen des sommes mobiles de naissances sur 12 mois selon le nombre p de mois de la droite d'ajustement et la distance h du mois sur lequel porte la prévision.
Source : G. Calot et R. Nadot. Combien y aura-t-il de naissances dans l'année ? Population, numéro spécial, 1977.

Dans ce sens nous envisageons, à la suite des travaux qui ont été effectués en France, de mettre rétrospectivement à l'épreuve diverses méthodes de prévisions. En comparant les résultats trouvés aux évolutions enregistrées, on est à même de porter un jugement sur la pertinence de ces méthodes.

Attachons-nous, tout d'abord, au problème de l'extrapolation des séries chronologiques de naissances. Sur le court terme, on travaillera sur des séries mensuelles. Ce pourra être :

- les séries de valeurs mensuelles désaisonnalisées;
- les sommes mobiles sur 12 mois.

Sur chacune de ces séries, on effectuera des extrapolations linéaires par la méthode des moindres carrés. La qualité de ces extrapolations sera appréciée à partir des valeurs des écarts quadratiques moyens entre réalisations et prévisions.

Les prévisions ainsi conduites pourront se distinguer selon :

- le nombre des derniers mois pris en compte, soit p ;
- l'importance de la période prospectée, soit h ;

la période passée utilisée étant la période 1960-1974.

Les résultats de cette confrontation entre prévisions rétrospectives et réalités, font l'objet des figures 3, 4, 5, et 6.

On parvient à des conclusions assez différentes selon que l'on utilise les naissances mensuelles des p derniers mois avant la date de point de départ de la prévision ou les séries des sommes de naissances arrêtées aux p derniers mois.

Avec les figures 3 et 4 on met en relief l'importance de la prise en compte du passé (p) selon l'ampleur de la prévision effectuée (h). En faisant intervenir les données mensuelles isolément (figure 3) et si l'on excepte la seule prise en compte de la donnée du dernier mois précédant le point de départ de la prévision (p =1) (prévoir revient alors à reconduire le dernier résultat observé), on voit que l'imprécision de la prévision est d'autant plus faible que l'on a utilisé les données d'une période antérieure plus étendue (toutefois, aux très longues périodes non représentées sur le graphique, l'imprécision augmente) ; en un mot, il convient d'utiliser un nombre de mois pas trop petit pour atténuer les effets sur l'ajustement des aléas de la série mensuelle. Avec l'utilisation des sommes mobiles sur 12 mois, on arrive à des profils d'erreurs, selon la durée du passé mise en jeu, très différents : rapidement cette durée entraîne une accentuation de l'erreur. Et si dans les deux cas, comme on pouvait s'y attendre, à prise en compte d'une durée passée égale l'erreur sur le futur est d'autant plus importante que la prospection est poussée plus loin (h croissant), la disparité selon la longueur de l'anticipation est plus grande quand on travaille sur des sommes mobiles (il convient pour effectuer cette comparaison de diviser les ordonnées du graphique de la figure 4 par 12).

Les figures 5 et 6 donnent une lecture différente des résultats précédents :

- sur la figure 5 si l'on voit que le gain en précision est nettement fonction de l'ampleur du passé mis à contribution pour extrapoler, le cas d'une simple reconduction de la valeur du dernier mois observé (p =1) correspond à la situation la plus avantageuse dès lors que la prospection porte sur 5 mois ou plus ;

- sur la figure 6, la situation est un peu plus complexe ; si toutefois l'importance du passé pris en compte est un facteur défavorable, le cas d'une simple reconduction de la somme mobile la plus récente (p =1) est cependant le choix le plus judicieux dès lors que la prospection porte sur 10 mois ou plus.

Si l'on veut maintenant chiffrer l'ordre de grandeur des erreurs résultant d'une telle conduite des opérations, on notera :

- qu'à la série mensuelle corrigée des variations saisonnières aux quelque 70 000 naissances mensuelles correspond une marge d'incertitude, définie par l'erreur quadratique moyenne, peu variable avec l'importance de l'horizon prospecté et de l'ordre de 1 500 à 2 500 naissances soit une erreur relative de l'ordre de 2 à 3,5 %.

- qu'à la série des sommes mobiles, aux quelque 840 000 naissances annuelles correspond une marge d'incertitude de 1 800 à 20 000 naissances selon que l'on anticipe de 1 mois ou de 12 mois, soit une erreur relative allant de 2 ‰ à 2,4 %.

Ainsi l'inertie de la série désaisonnalisée apparaît beaucoup plus grande que la série des sommes mobiles et l'on vérifie qu'une prévision à 1 mois est meilleure à partir de la première (on adopte alors selon la figure 5, p = 1) que celle conduite à partir de la seconde (en faisant choix de p = 2, ainsi que le suggère la figure 6).

Tableau 2 - Caractéristiques de la primo-nuptialité de quelques générations masculines françaises.

Générations	Répartition de 1000 premiers mariages							Premiers mariages conclus avant 50 ans	
	18-20 ans	20-25 ans	25-30 ans	30-35 ans	35-40 ans	40-45 ans	45-50 ans	Nombre pour 1000 célibataires à 18 ans	âge moyen
1821-1825	22	253	365	225	79	39	17	890	28.7 yrs
1871-1875	28	317	389	172	55	28	11	900	28.0 yrs
1906-1910	39	439	355	89	39	28	11	900	26.3 yrs
1926-1930	38	470	339	88	38	16	11	915	25.9 yrs

Source: Jean-Claude Chasteland and Roland Pressat. La nuptualité des générations françaises depuis un siècle. Population, 1962, no. 2

Séries chronologiques et modèles mathématiques.

Depuis plus ou moins longtemps on a cherché à représenter par des fonctions mathématiques le déroulement dans le temps des phénomènes démographiques, les premières tentatives ayant concerné la mortalité (lois de Gompertz et de Gompertz-Makeham).

Les essais de formalisation concernant les autres phénomènes, la fécondité essentiellement, ont vu le jour beaucoup plus tardivement, ainsi en 1931, quand Wicksell y a vu un moyen de résoudre l'équation de Lotka qui commande l'établissement de la population stable limite impliquée par le maintien de conditions stationnaires de mortalité et de fécondité. A cette époque, la référence, en tant que moyen de vérifier la bonne adéquation du modèle, était fournie par la série, une année donnée (ou une époque de plusieurs années) des taux de fécondité par âge.

En quoi l'existence de séries chronologiques, en l'occurrence de séries de taux de fécondité par âge, peut-elle être mise à contribution pour la détermination d'une fonction fécondité représentative de la fécondité de la population en cause ? Pour répondre à cette question, faisons quelques remarques sur la nature des fonctions mathématiques les plus aptes à représenter la fécondité par âge.

On peut retenir à cet égard trois types de densité pour ces fonctions, toutes les trois appartenant à la famille du système de K. Pearson :

- la densité de probabilité attachée à la fonction gamma, cas particulier du type III de Pearson ;
- la densité de probabilité attachée à la fonction beta, cas particulier du type I de Pearson ;
- la forme polynomiale $(x-a)(\beta(x)^2$, cas particulier également du type I de Pearson.

Dans ces trois expressions les mieux adaptées à la représentation des fonctions fécondité, la forme mathématique de la densité de fécondité peut s'écrire :

$$f(x):=D_\beta\gamma(x)$$

ou $D_\beta$est la descendance finale (ou l'indice synthétique de fécondité si l'on est en transversal), $\gamma(x)$ étant le calendrier de la fécondité $(\int_\alpha^\beta\gamma(x)dx=1)$. Il apparaît donc, d'après la structure de cette formule, que nous ne savons pas déboucher sur une expression mathématique de la densité de fécondité qui exprimerait la solidarité entre l'intensité et le calendrier du phénomène.

Séries chronologiques et projections.

"En l'absence de lois permettant de prévoir à coup sûr, l'étude des liaisons entre situations à diverses dates plus ou moins espacées est le seul fondement scientifique de la prévision". "Prévoir le lendemain consiste à passer d'aujourd'hui à demain connaissant hier". Ces deux citations extraites de l'étude de L. Henry et H. Gutierrez ("Qualité des prévisions démographiques à court terme". Population n° 3, 1977), précisent bien le caractère des prévisions démographiques et la place que peut tenir l'analyse des séries chronologiques dans l'établissement des dites prévisions.

Distinguons le très court terme d'une part et le moyen et long terme d'autre part.

Les prévisions à très court terme.

Leur objectif est de déterminer le sens exact des évolutions les plus récentes avec l'espoir de mettre en évidence d'éventuels retournements de tendance.

Illustrons le présent propos en rapportant la série des proportions de célibataires masculins découlant de la longue suite des recensements quinquennaux français allant de 1851 à 1946 (en 1956, 1961 et 1966, des estimations suppléent les recensements manquants à ces dates) (tableau 1). En s'en tenant à quelques générations types, le lecteur pourra tirer des données de ce tableau, les conclusions suivantes (tableau 2).

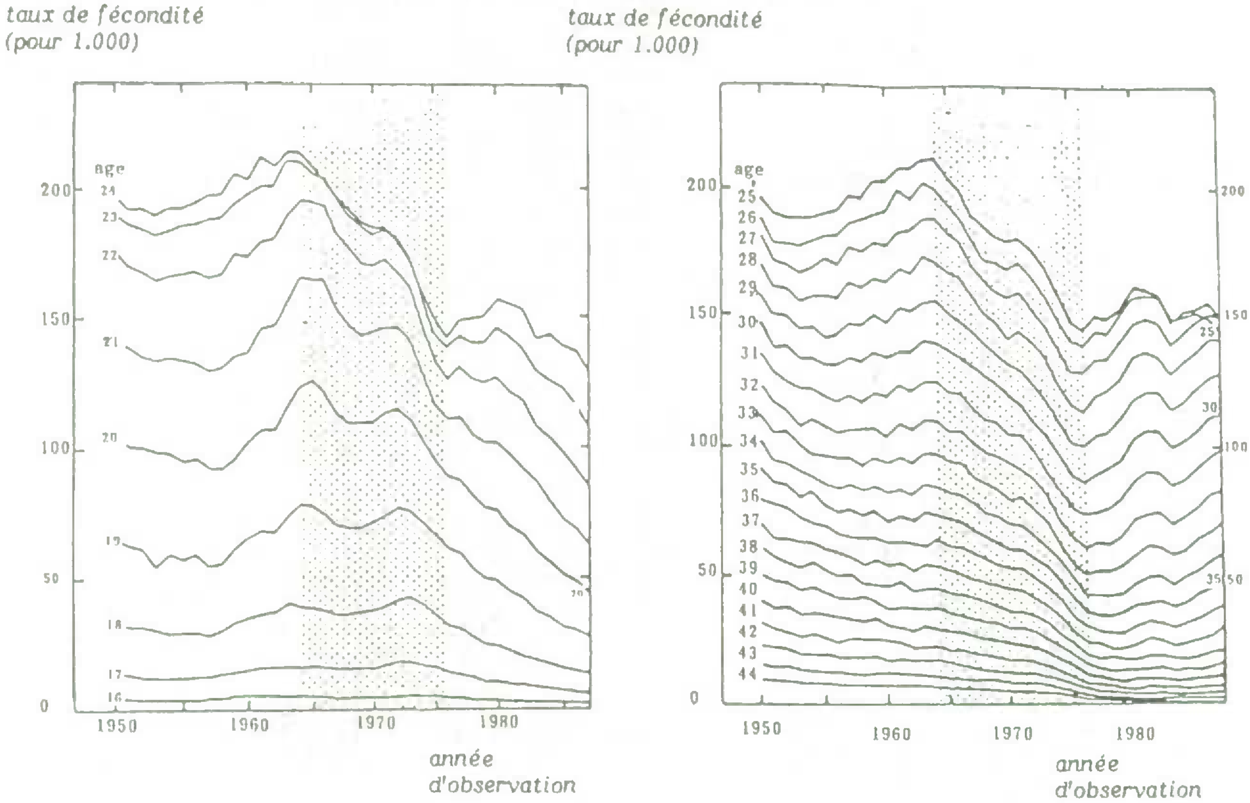


Figure 2 - France. Evolution des taux de fécondité générale par âge.
Source : G. Calot. Symposium on Population Change and European Society. European University Institute. Florence 7-10 décembre 1988.

Tableau 1 - France. Proportions (en %) d'hommes célibataires dans différents groupes de générations.

Groupe d'âges		Générations												
		1821 1825	1826 1830	1831 1835	1836 1840	1841 1845	1846 1850	1851 1855	1856 1860	1861 1865	1866 1870	1871 1875	1876 1880	1881 1885
15-19	ans	(99.7)	(99.7)	99.8	99.7	99.7	99.2	99.7	99.9	99.4	99.8	99.9	99.8	99.7
20-24	ans	(89.0)	89.4	80.2	87.3	78.9	82.6	84.4	86.8	86.8	90.4	92.7	90.4	(87.0)
25-29	ans	58.3	54.8	56.0	48.7	49.0	45.4	48.6	50.2	50.5	48.9	48.1	(46.0)	43.6
30-34	ans	31.0	30.1	26.9	28.3	25.8	28.8	29.6	27.4	26.6	23.6	24.2	22.8	(26.0)
35-39	ans	19.0	17.8	18.8	18.1	20.2	21.0	18.9	18.0	16.3	(16.8)	15.7	(16.5)	(15.6)
40-44	ans	13.5	17.5	14.0	15.7	16.8	15.3	14.5	(12.7)	(12.9)	12.6	(12.8)	(12.4)	9.9
45-49	ans	12.0	11.9	12.9	14.7	13.1	12.8	10.9	(11.3)	11.3	(11.3)	(10.6)	10.6	9.3

Groupe d'âges		Générations											
		1886 1890	1891 1895	1896 1900	1901 1905	1906 1910	1911 1915	1916 1920	1921 1925	1926 1930	1931 1935	1936 1940	1941 1945
15-19	ans	(99.7)	(99.7)	(99.9)	99.4	99.4	99.4	99.5	(99.5)	99.2	99.4	99.4	99.3
20-24	ans	(89.3)	(90.0)	(80.0)	76.4	78.0	79.3	(87.0)	80.6	76.9	77.4	78.5	-
25-29	ans	(54.0)	(40.5)	34.7	36.0	35.7	(42.0)	46.0	36.2	32.6	34.5	-	-
30-34	ans	(23.7)	20.9	18.7	20.2	(20.0)	26.1	21.0	18.1	17.6	-	-	-
35-39	ans	10.6	12.7	13.2	(15.0)	16.8	15.6	12.9	13.1	(12.8)	-	-	-
40-44	ans	10.2	10.3	(10.5)	12.5	13.0	11.0	10.5	(10.5)	(10.3)	-	-	-
45-49	ans	8.7	(9.0)	9.6	10.8	10.7	10.1	(9.4)	(9.4)	(9.0)	-	-	-

Nota : Les pourcentages entre parenthèses correspondent à des estimations

Source: J.-C. Chasteland et R. Pressat. La nuptialité des générations françaises depuis un siècle. Population, n° 2, 1962.

ou une croissance (en d.) là où respectivement croissance et décroissance avec l'âge persistent tandis qu'en b. et c. la décroissance et la croissance avec l'âge qui sont observées dans les générations sont amplifiées dans une vision transversale (on notera, toutefois, qu'à côté des schémas précédents, d'autres typologies peuvent exister qui entraînent des distorsions moins importantes).

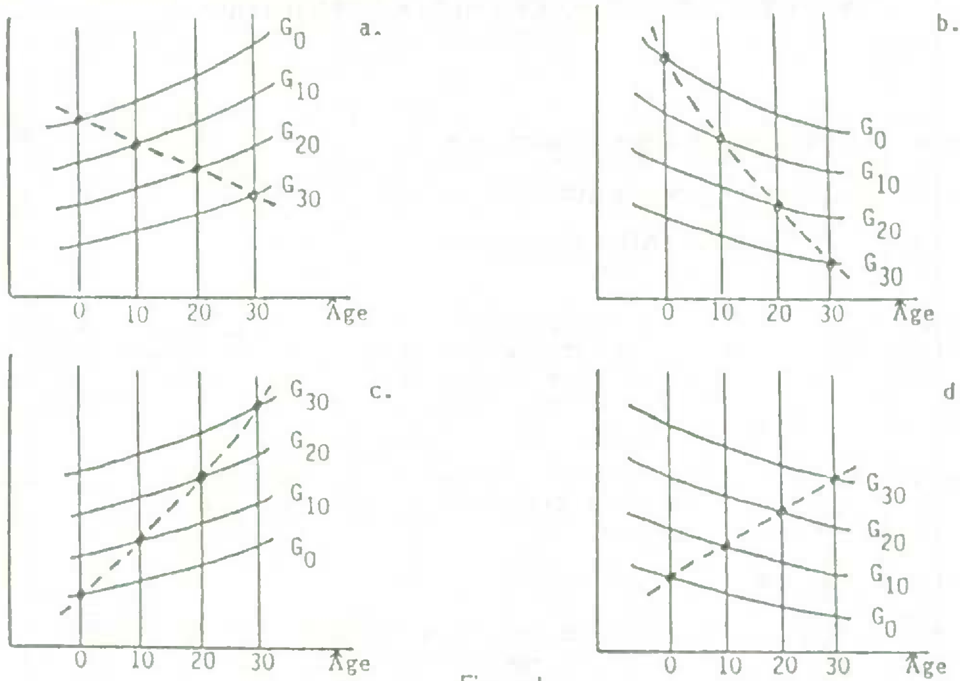


Figure 1

- En matière de fécondité, on accorde d'emblée la priorité à une vision longitudinale du phénomène en raison du poids que peut avoir le passé dans le comportement présent des générations. A défaut de procéder à des reconstitutions de l'histoire des générations fécondes, on peut, et parfois avantageusement, procéder à une lecture de la suite des données annuelles qui prennent en compte les péripéties passées de la vie génésique des femmes. Avec la figure 2, où sont rapportés les taux de fécondité générale par âge en France de 1950 à 1987, on a une illustration exemplaire de ce mode de lecture. Limitons notre commentaire à la période postérieure à 1964, date à partir de laquelle la fécondité a décru très sensiblement. Cette période de grande baisse se situe de 1964 à 1976 (dans cet intervalle, l'indice synthétique de fécondité est passé de 2,90 naissances par femme à 1,83). La baisse s'observe sans répit à tous les âges après 28 ans ; avant 29 ans et vers le milieu de la période, on observe une pause et, parfois, une reprise d'ampleur variable. Après 1976, la baisse est soutenue avant 22 ans, hésitante dans un premier temps de 22 à 25 ans, elle manifeste une tendance à la reprise au-delà avec toutefois un reflux brutal et généralisé en 1983. On est naturellement conduit à s'interroger sur le comportement contrasté des jeunes générations et des générations plus âgées et c'est là qu'il convient d'invoquer le passé des unes et des autres : la remontée récente de la fécondité aux âges allant d'environ 27 ans jusque vers 37 ans tient, à n'en pas douter, à une modification de calendrier marqué par des naissances beaucoup plus rares en début de période fertile avec, comme corollaire, un léger regain en fin de période. Cette analyse explique en même temps qu'elle contient pour le futur la promesse d'une stabilisation de l'indice synthétique de fécondité au cours des prochaines années, dans la mesure où le bas niveau actuel de fécondité des plus jeunes générations finissant par se stabiliser, la récupération partielle en fin d'histoire génésique assurera une compensation au cours de cette période transitoire précédant l'installation d'un éventuel régime stationnaire.

Séries chronologiques d'états de population.

De telles séries se fondent sur une suite de recensements et elles présentent un intérêt particulier quand la périodicité de ces recensements est bien régulière, les intervalles intercensitaires étant égaux à l'étendue des classes d'âges. Pendant longtemps les recensements français ont été exemplaires à cet égard, leur périodicité quinquennale s'ajustant à la décomposition de la population par groupes d'âges de cinq ans ; dans ces conditions, on dispose d'un bon suivi, au fil du temps, de l'évolution des différents groupes de cinq générations. Hajnal (et avant lui G. Mortara, mais de façon moins systématique) a été le premier à montrer pleinement l'intérêt que représente la donnée des proportions de célibataires, par sexe et âge, en tant que moyen d'étudier la primo-nuptialité. Ces taux que l'on peut appeler taux de célibat ont leur équivalent formel dans les taux d'activité, les taux de diplômés d'un certain degré, les taux de fréquentation scolaire,...Ils représentent, à de légers facteurs correctifs près, les célibataires de la table de primo-nuptialité, les actifs de la table d'activité,... En matière de fécondité, le renseignement qui apparaît est l'étendue moyenne de la descendance déjà constituée (soit que l'on connaisse le détail des diverses dimensions de famille déjà atteintes, soit que l'on ne dispose que de la dimension moyenne) ; le renseignement s'inscrit ici dans la série des événements cumulés de la table de fécondité.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

LES SÉRIES CHRONOLOGIQUES EN DÉMOGRAPHIE

R. Pressat[1]

Eléments de base des séries chronologiques en démographie

A la base des diverses séries chronologiques en démographie, on trouve :

- des séries d'événements : naissances, mariages, divorces, décès,...
- des séries d'états de population.

Ces séries peuvent correspondre à des découpages de temps variables. Si les événements sont le plus souvent comptabilisés par années, ils peuvent l'être, faute de mieux, par périodes pluri-annuelles ou encore, spécialement en vue d'études spécifiques, sur des courtes durées (le mois, le jour, voire l'heure). Quant aux états de population, leur périodicité est liée le plus souvent à celle des recensements ou aux estimations effectuées par les offices de statistique au cours des périodes intercensitaires.

Signalons, pour l'écarter dans la suite de notre propos, le cas où les événements sont comptabilisés sur de petits intervalles afin de faire apparaître sur une période déterminée (l'année, la semaine, le jour) comment se trouve rythmée la venue du type d'événement considéré.

Nature des séries chronologiques

Les séries de données brutes précédentes ne sauraient donner lieu qu'à des interprétations limitées ; elles constitueront, le plus souvent, la matière première permettant d'établir des séries chronologiques d'indices se prêtant mieux à l'analyse. Au nombre de ces derniers, citons ceux résultant du calcul de rapports ou proportions :

- les taux bruts (essentiellement de natalité et mortalité) ;
- les taux spécifiques du type taux par âge ou, plus généralement, par durée (exemple : taux de fécondité par durée de mariage).
- les quotients d'éventualité tels que les quotients de mortalité, de primo-nuptialité,...
- les taux en tant que rapports de deux effectifs de population, celui figurant au numérateur se rapportant à une sous-population appartenant à la population retenue au dénominateur (ainsi les proportions de célibataires par sexe et âge à une date donnée, les taux d'activité, de scolarité,...).

Les séries chronologiques, base de l'analyse démographique

Les phénomènes démographiques, comme tous les phénomènes sociaux, sont immergés dans le temps et l'on ne saurait en conduire l'analyse sans se référer à cette dernière dimension. En bref, la constitution et l'analyse des séries chronologiques sont au coeur de l'analyse démographique.

Ces séries peuvent être analysées pour elles-mêmes ou en conjonction avec d'autres séries de manière à dégager des associations entre les modalités de survenance des phénomènes démographiques et celles d'autres phénomènes relevant de la vie des individus en société.

Un premier dilemme : analyse transversale ou analyse longitudinale ?

En présence d'une série de taux (par âge pour fixer les idées) ou d'indices résumés (cumuls de taux de fécondité générale, vie moyenne), comment doit-on mener l'analyse ?

- En matière de mortalité, il est d'usage de reconnaître la priorité à l'analyse transversale. A cela deux raisons : le souci de suivre l'actualité au plus près et ainsi de reconnaître toute évolution fâcheuse afin d'y remédier et, plus fondamentalement, la conviction que la situation du moment se ressent peu ou pas de la situation passée. Toutefois, cette attitude n'exclut pas qu'en présence de phénomènes à évolution rapide des synthèses longitudinales aient leur place, la vision transversale pouvant se révéler déformante. Ce type de distorsion relève d'un schéma très général synthétisé par la figure 1. Les lignes en traits pleins sont affectées à des générations distantes de 10 années G_0 ,G_{10},, G_{20}, G_{30}, et représentent la variation avec l'âge du risque attaché à un phénomène. En a. et b. ce risque, à âge égal, croît à mesure qu'il s'agit d'une génération plus récente mais alors qu'en a. il croît avec l'âge dans une même génération, en b. il décroît ; en c. et d. le risque, à âge égal, décroît à mesure qu'il s'agit d'une génération plus récente, mais alors qu'en c.il croît avec l'âge dans une même génération, en d. il décroît. La synthèse transversale obtenue par emprunt de points convenables aux lignes de générations (lignes en pointillées), donne une vision tout autre du phénomène suggérant une décroissance (en a.)

[1] R. Pressat, Département de la Conjoncture, Institut national d'études démographiques, Paris, France.

de l'état matrimonial). De plus, la microsimulation est la seule façon directe d'évaluer, par exemple, la proportion de mariages qui se terminent par un divorce ...dans le cadre des modèles.

NOTES

1. Comme le caractère permanent d'une séparation (c.-à-d. une séparation sans réconciliation subséquente) ne peut pas être confirmé avant le divorce (ou le décès d'un des conjoints), les données sur les séparations surestiment les ruptures matrimoniales. De même, certains couples séparés peuvent ne jamais demander le divorce pour des motifs religieux ou autres, ce qui fait que les données sur le divorce sous-estiment les ruptures matrimoniales. Néanmoins, la date de la séparation donne une idée plus juste du moment où se produit une rupture matrimoniale que la date du divorce.

2. L'utilisation de la variable du niveau de scolarité atteint au moment de l'enquête plutôt qu'au moment du mariage n'entraîne pas simplement une certaine confusion quant à la période de référence. Elle donne lieu à un biais de choix parce que les années-personnes pour lesquelles il y a un risque de mariage à des niveaux de scolarité moins élevés n'incluent pas les premières années d'études des personnes qui atteignent à la longue un niveau de scolarité plus élevé. Le biais gonflerait généralement les taux de mariage des personnes qui ont un niveau de scolarité peu élevé. Des variables telles que la pratique religieuse (au moment de l'enquête), qui sont sujettes à changer, peuvent aussi donner lieu à des biais de choix semblables.

3. Les transitions des états initiaux CEL, ULI ou SEP tendent à se produire dans un intervalle relativement court. Par conséquent, dans ces cas, il a été possible de se limiter aux années-personnes récentes à risque tout en conservant un échantillon de taille adéquate. La date à partir de laquelle la réforme de la Loi sur le divorce est entrée en vigueur (juillet 1968) constitue un point repère pratique permettant de déterminer ce qui peut être considéré comme récent. Toutefois, étant donné que les mariages ont tendance à être de longue durée, toutes les années-personnes possibles présentant un risque de séparation ont été utilisées (c.-à-d. pour les transitions MAR SEP). Cette façon de procéder s'est avérée particulièrement importante pour l'évaluation des risques relatifs d'une séparation après le départ du foyer du dernier enfant.

BIBLIOGRAPHIE

Balakrishnan, T.R., Rao, K. Vaninadha, Lapierre-Adamcyk, Evelyne, and Krotki, Karol J. (1987), "A Hazard Model Analysis of the Covariates of Marriage Dissolution in Canada, "*Demography*, 24(3), 395-406.

Burch, Thomas K. (1985), "*Family History Survey: Preliminary Findings*, Catalogue 99-955, Statistics Canada, Ottawa.

Burch, Thomas K. and Madan, Ashok K. (1986), Union Formation and Dissolution: Results from the 1984 Family History Survey, Catalogue 99-963, Statistics Canada, Ottawa.

Cox, D.R., and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.

Grenier, Gilles, Bloom, David E., and Howland, D. Juliet (1987), "An Analysis of First Marriage Patterns of Canadian Women", *Canadian Studies in Population*, 14(1), 47-68.

Hannan, Michael T., and Tuma, Nancy Brandon (1978), "Income and Independence Effects on Marital Dissolution: Results from the Seattle and Denver Income-Maintenance Experiments", *American Journal of Sociology*, 84(3), 611-633.

Hoem, Britta, and Hoem Jan M. (1988), "Dissolution in Sweden: The break-up of conjugal unions to Swedish women born 1936-60", Stockholm Research Reports in Demography 45, University of Stockholm, Section of Demography.

Hoem, Jan M. (1985), "Weighting misclassification, and other issues in the analysis of survey samples of life histories", in *Longitudinal Analysis of Labor Market Data*, eds. James J. Heckman and Burton Singer, Cambridge: Cambridge University Press, 249-293.

__________ (1989), "Limitations of a Heteregeneity Technique: Selectivity Issues in Conjugal Union Disruption at Parity Zero in Contemporary Sweden", Stockholm Research Reports in Demography 56, University of Stockholm, Section of Demography.

Hogan, D. (1978), "The Effects of demographic factors, family background and job achievement on age at marriage", *Demography*, 15(1), 155-175.

McCullagh, P., and Nelder, J.A. (1983), *Generalized Linear Models*, London: Chapman and Hall.

Teachman, Jay D., Polonko, Karen A., and Scanzoni, John (1987), "Demography of the Family", in *Handbook of Marriage and the Family*, eds. Marvin B. Sussman and Suzanne K. Steinmetz, New York: Plenum Press, 3-36.

Vaupel, James W., and Yashin, Anatoli I. (1985), *Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics*, American Statistician, 39(3), 176-185.

TABLEAU 4
Risques relatifs d'une séparation selon l'activité et la présence d'enfants

Présence d'enfants				
Sans enfants	1			
Enfants d'âge préscolaire (1 +, ≤ 6 ans)	0.78			
Enfants d'âge scolaire (tous de $\overline{7 +}$)	1.19			
Nid vide	5.87			
Situation relative à l'emploi	Hommes	Femmes		
Actif	1	1.86		
Inactif	1.63	0.68		
Jamais travaillé	0.68	0.81		
Commence à travaillé	0.94	2.09		
Arrête de travaillé	2.75	1.28		
Pas de réponse	0.79	0.83		
Antécédents professionnels (depuis le mariage)	Sans enfants		1 + enfants nés	
	Hommes	Femmes	Hommes	Femmes
Pas d'interruption du travail	1	0.57	1	1
Interruption du travail	1.10	0.68	0.61	0.76
Sans travail depuis le mariage	1.74	0.54	1	1
Coefficients de durée (pentes de risques logarithmiques)				
Durée de l'union (année de mariage + d'ULI)	-0.1065			
Durée de l'union sans efants (0 après la 1re naissance	0.1022			
Âge de l'enfant le plus vieux à la maison (log)	0.4508			
Âge de l'enfant le plus jeune à la maison (log)	-0.1760			

Le tableau 4 indique les risques relatifs découlant de l'interaction des variables de la situation relative à l'emploi et de la présence d'enfants dans la régression qui réalise l'ajustement optimal. La constatation la plus notable est le lien qui existe entre les régimes d'activité non traditionnels (par ex., l'homme qui ne travaille pas, ou la femme qui travaille, particulièrement s'il y a des enfants à la maison) et les risques élevés de séparation des couples mariés.

Les résultats de la figure 4 et du tableau 4 ne sont pas compatibles avec ni l'une ni l'autre des théories selon lesquelles il est normal que les mariages se "détériorent" ou qu'un grand nombre de mariages sont instables au départ. Plutôt, les risques d'une séparation diminuent normalement avec le temps, mais le nombre et l'âge des enfants au foyer exercent aussi un effet sur ces risques. Il n'est pas nécessaire d'invoquer l'hétérogénéité non observée pour expliquer les épisodes de risque élevé qui peuvent être reliés, par exemple, au fait que le plus jeune enfant atteint l'âge scolaire et (ou) que la mère réintègre la population active.

5. DISCUSSION

Deux exemples d'analyse des données de l'EF ont été fournis dans le présent rapport. Dans un cas, il a été démontré que des changements relatifs à la situation vis-à-vis de l'activité influent sur les taux de mariage et d'ULI. Implicitement, les facteurs qui ont une incidence sur le marché du travail permettraient d'expliquer en partie l'évolution des taux de mariage et les tendances relatives à l'âge moyen au mariage. De même, un lien a été établi entre le risque de séparation d'une part, et la fécondité matrimoniale/situation vis-à-vis de l'activité d'autre part. Par conséquent, l'évolution des taux de divorce reflétera en partie les régimes de fécondité et d'activité des hommes et des femmes.

On a laissé entendre que la dynamique de la prise de décisions relatives à l'état matrimonial, à la naissance d'enfants et à l'activité doit être examinée très attentivement au moment de la modélisation des taux de mariage pour que nos modèles donnent une idée juste du comportement des couples, et cette implication est plus importante que les exemples précis qui ont été fournis. Le fait de se concentrer sur la dynamique permet aussi de déterminer les sujets qui peuvent ou non être traités directement à l'aide des données disponibles. Par exemple, les résultats de la section 3.2 démontrent clairement qu'il est extrêmement difficile de tenter de déterminer la proportion de mariages qui aboutiront au divorce. La source de la difficulté est que le risque de divorce dépend des régimes d'activité dynamiques comportant des interruptions du travail et du nombre d'enfants de même que des intervalles de naissance.

Les régressions qui ont été utilisées dans le présent rapport ont été intégrées à un modèle de microsimulation démographique (DEMOGEN) qui comprend des modèles de l'état matrimonial et de l'activité établis à partir des données de l'EF. La création d'un modèle de microsimulation semble être la seule solution permettant d'intégrer et d'évaluer les implications des modèles de chronologie des événements complexes (par ex., les transitions matrimoniales dépendant de la situation vis-à-vis de l'activité et les transitions relatives à l'activité,

4.2 Séparation, activité et âge des enfants

L'idée voulant que l'on ait envie de changement après un certain nombre d'années de vie commune ("seven year itch") fait partie du folklore nord-américain tout comme la notion que certains mariages durent uniquement à cause des enfants (surtout si ces derniers sont jeunes). Le fondement empirique de la première croyance populaire repose sur la constatation que les taux de divorce observés selon la durée du mariage augmentent souvent durant les premières années de l'union, puis décroissent après de plus longues périodes, ce qui laisse supposer, dans un certain sens, que les mariages se "détériorent" à la longue.

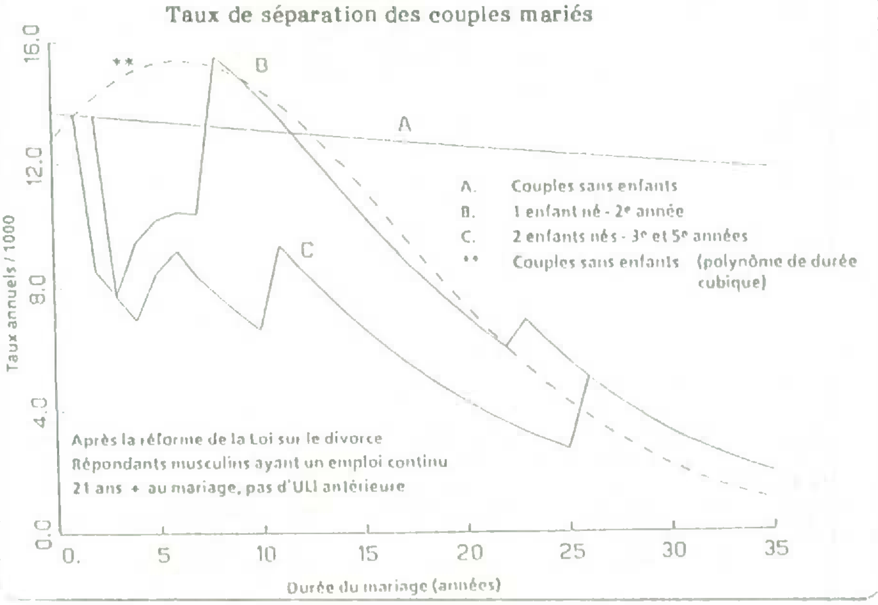
D'après une autre explication des tendances du divorce en fonction de la durée du mariage, les mariages pourraient dès le départ être répartis en deux catégories, soit les mariages stables et les mariages instables. Les mariages stables présentent des risques fixes et peu élevés de divorce. Les mariages instables comportent des risques qui augmentent régulièrement avec le temps (c.-à-d. une fois la lune de miel terminée). Etant donné que de façon générale, il n'est pas possible de faire de telles distinctions, nous obtenons un taux de divorce qui représente la moyenne de deux taux, mais sur lequel le taux de divorce des mariages instables exerce un plus grand effet (c.-à-d. que ce taux fait grimper le taux moyen de divorce). En effet, lorsqu'on ne tient pas compte des mariages instables, le taux de divorce des mariages stables influe davantage sur le taux moyen de divorce qui baisse à un niveau relativement faible. Ainsi, la structure complexe de durée pourrait tout aussi bien être expliquée par une combinaison de structures moins complexes (c.-à-d. les effets de l'hétérogénéité non observée, Vaupel et Yashin, 1985).

La figure 4 montre quelques-uns des résultats de l'analyse des données de l'EF sur la séparation. La ligne pointillée (caractérisée par **) est fondée sur une régression qui comprend un polynôme cubique de la durée des mariages pour représenter une tendance à la hausse puis à la baisse de la structure de durée du risque de séparation (que l'on retrouve chez les mariages sans enfants aussi bien que les mariages avec enfants).

Les trois autres lignes (désignées par les lettres A, B et C correspondant respectivement aux couples sans enfants, avec un enfant et avec deux enfants) sont fondées sur l'équation la mieux ajustée qui comporte des variables de durée multiples: la durée de l'union (années vécues en ULI + années de mariage), la durée de l'union sans enfants (0 après la première naissance), l'âge de l'enfant le plus vieux à la maison et celui du plus jeune (peut être le même). Jusqu'à présent, une seule autre étude semble avoir porté sur des interactions de fécondité semblables avec la variable de durée (Hoem et Hoem, 1988).

La complexité de la figure 4 est due au fait qu'une ligne représentant les risques de séparation selon la durée du mariage pour une chronologie donnée d'événements relatifs à l'état matrimonial, à l'activité et à la fécondité peut changer soudainement. La naissance d'un enfant additionnel pourrait être la cause d'un tel changement. L'importance du changement dépendrait du nombre de naissances précédentes et de l'intervalle entre ces naissances. On relève aussi de brusques changements lorsqu'il n'y a plus d'enfants d'âge préscolaire ou lorsque le dernier enfant quitte le foyer (le syndrome du nid vide). La figure 4 contient deux lignes qui représentent les mariages sans enfants. Il est toutefois très clair que la ligne pointillée (**) correspond à l'amalgame des effets de la durée sur le risque de séparation avec la manifestation dans le temps de la fécondité matrimoniale. L'équation la mieux ajustée sous-entend que les risques de séparation diminuent généralement avec la durée (même dans le cas des couples sans enfants de la ligne A), bien que les risques puissent augmenter à mesure que les enfants atteignent des âges critiques. La présence d'enfants est un facteur qui contribue à la fois à l'augmentation et à la diminution des risques de séparation.

FIGURE 4
Taux de séparation des couples mariés



Après la réforme de la Loi sur le divorce
Répondants musculins ayant un emploi continu
21 ans + au mariage, pas d'ULI antérieure

les années de scolarité terminées. Ainsi, les données sur l'âge au premier emploi et les années subséquentes d'activité peuvent être utilisées comme substituts pour la scolarité et l'expérience professionnelle (les variables relatives aux ressources humaines sont souvent employées par les économistes pour analyser les différences entre les taux de rémunération).

FIGURE 3
RISQUES RELATIFS DU MARIAGE

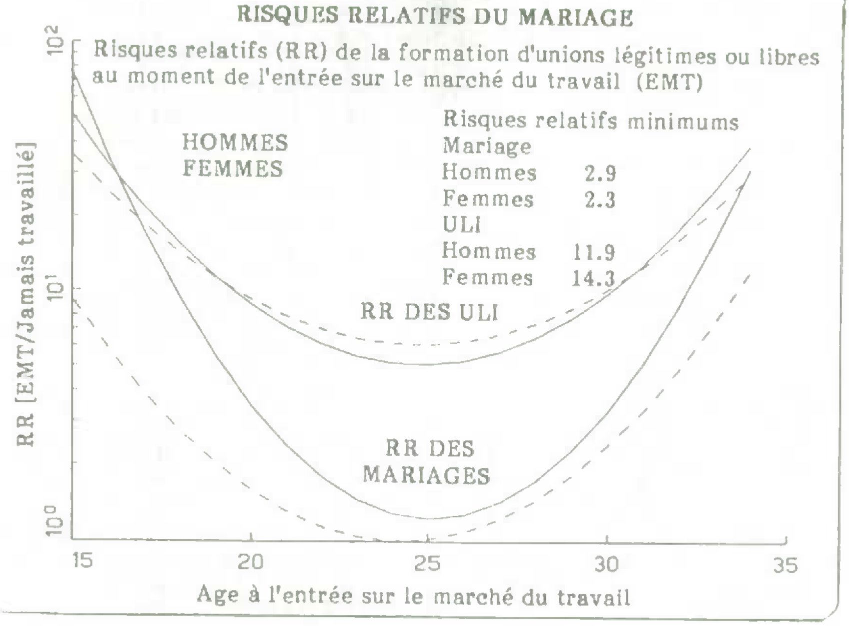


TABLEAU 3
Risques relatifs de la formation d'unions légitimes ou libres
selon la situation relative à l'emploi

Antécédants de travail	Années de travail complètes							
Âge à l'entrée sur le marché du travail	0	5	10	15	0	5	10	15
HOMMES	Mariage				ULI			
15	0.70	0.71	0.71	0.72	0.91	0.75	0.61	0.50
20	1	0.76	0.58	0.44	1	0.64	0.41	0.27
25	1.42	0.82	0.48	0.28	1.10	0.56	0.28	0.14
30	2.03	0.89	0.39	0.17	1.22	0.48	0.19	0.07
FEMMES								
15	0.90	0.77	0.66	0.57	1.15	0.84	0.62	0.45
20	1	0.60	0.36	0.22	1	0.50	0.24	0.12
25	1.11	0.47	0.20	0.09	0.87	0.29	0.10	0.03
30	1.,23	0.37	0.11	0.03	0.75	0.17	0.04	0.01
Situation relative à l'emploi	Hommes				Femmes			
Actif		1	1			1	1	
Inactif		0.26	0.71			0.80	0.70	
Commence à travailler		0.45	0.56			0.51	0.61	
Arrête de travailler		0.57	0.60			2.22	1.60	
Retour aux études		0.47	1.12			1.06	0.81	

L'observation selon laquelle la possibilité de se marier augmente typiquement avec l'âge à l'entrée sur le marché du travail, mais décroît avec les années d'activité laisse supposer qu'il n'y a peut-être pas de rapport simple (s'il y en a effectivement un) entre le revenu et le moment choisi pour le mariage, étant donné qu'il devrait exister un lien positif entre la scolarité, l'expérience professionnelle et le revenu. Néanmoins, le risque relatif se rapportant à la catégorie "Inactif" (tableau 3) indique que les interruptions du travail entraînent une diminution de la possibilité de se marier (diminution plus marquée pour les hommes que pour les femmes). A cet égard, les facteurs économiques peuvent avoir une incidence directe. Signalons que la catégorie "Inactif" se distingue des catégories de situations indéterminées ("Commence à travailler" et "Arrête de travailler") et n'inclut pas les périodes où le retour aux études constituait la raison de l'interruption du travail.

de la population active atteindra généralement. La seule réserve importante à cet égard serait une grossesse qui aurait un effet l'emportant sur tous les autres effets pour la durée de l'événement. Il est douteux que l'effet d'indépendance puisse représenter entièrement une relation causale. Il peut souvent arriver que l'entrée sur le marché du travail soit motivé par la décision de se marier.

TABLEAU 2
Taux de transition annuels / 1000 à partir d'équations de régression

ÉTAT INITIAL		ÉVÈNEMENT	
CÉL			
Jamais travaillé			
20 ans, pas de grossesse		Mariage	Formation d'une ULI
Pas d'ULI antérieure	Homme	1.8	1.1
	Femme	4.7	2.0
Entrée sur le marché du travail			
20 ans, pas de grossesse			
Pas d'ULI antérieure	Homme	14.8	22.3
	Femme	17.4	42.6
ULI			
Entrée sur le marché du travail			
1re année ULI			
20 ans, pas de grossesse		Mariage	Dissolution de l'ULI
Pas d'ULI antérieure	Homme	19.7	19.8
	Femme	23.2	15.7
MAR			
Après la réforme de la Loi sur le divorce			
Sans enfants, marié à 21 ans +			
Pas d'ULI antérieure			
1re année de mariage			Séparation
Pas d'interruption du travail	Homme		13.7
Pas travaillé depuis le mariage	Femme		6.5
SEP			
2^{e} année de séparation			Divorce
Enfants d'âge préscolaire à la séparation			
Pas de nouvelle ULI			272.2

Une manifestation de l'effet d'indépendance peut être observée dans les taux de transition annuels de l'état de CEL (tableau 2) qui indiquent une différence marquée entre les répondants qui n'ont jamais travaillé et ceux qui sont entrés sur le marché du travail à 20 ans. Les équations de régression indiquent le même genre de résultats pour les répondants qui avaient entre 15 et 35 ans au moment de leur intégration dans la population active. La figure 3 compare les risques au moment de l'entrée sur le marché du travail aux risques avant le début de l'activité. L'axe vertical est le risque relatif calculé à partir du risque relatif établi pour les nouveaux actifs divisé par le risque associé aux répondants du même âge qui n'ont jamais travaillé. Il est clair que pour les répondants de plus et de moins de 20 ans, l'effet d'indépendance peut être encore plus grand que ne l'indiquent les données du tableau 2.

Le tableau 3 présente les risques relatifs pour les célibataires de la formation d'une union légitime ou libre durant les années suivant leur entrée sur le marché du travail. Pour chaque âge à l'entrée, la possibilité de contracter un mariage ou de vivre en union libre diminue régulièrement par la suite. Toutes choses étant égales par ailleurs, les taux de mariage atteignent un sommet au cours des années qui suivent immédiatement l'entrée sur le marché du travail. Toutefois, les données chronologiques sur la situation relative à l'emploi de l'EF ont été déclarées en années plutôt qu'en années et en mois. L'effet de l'arrondissement à l'année la plus proche est qu'il y a des mois où la situation relative à l'emploi est imprécise (catégories **"commence à travailler"** et **"arrête de travailler"**) et pour lesquels il est impossible d'établir avec précision l'ordre des événements (par ex., si l'entrée sur le marché du travail a eu lieu avant le mariage). Ainsi, on ne peut pas déterminer si l'effet d'indépendance est relié à une brève période qui suit immédiatement l'entrée sur le marché du travail plutôt qu'à une période couvrant l'année ou les deux ans qui suivent cette intégration.

L'analyse des données pour l'étape qui suit l'entrée sur le marché du travail a donné lieu à une équation réalisant l'ajustement optimal et incluant l'âge à l'entrée et les années d'activité utilisés comme variables distinctes avec des coefficients de signes différents (sauf dans le cas de la formation d'ULI chez les femmes). Il importe de signaler que ces variables ont des effets différents étant donné que leur somme correspond à l'âge actuel dans la plupart des cas (sauf lorsqu'il y a eu des périodes d'interruption du travail). Cela sert à indiquer que les taux de mariage désagrégés selon le groupe d'âge seulement (c.-à-d. les cohortes des naissances) masqueront une hétérogénéité importante parmi les cohortes de la population active.

Des données chronologiques sur la scolarité n'ont pas été recueillies dans le cadre de l'EF; cependant, la question sur la date du premier emploi précisait qu'il ne fallait pas tenir compte des emplois occupés par les étudiants à plein temps. Par conséquent, l'âge au premier emploi peut être considéré comme un substitut pour

$$f(T|X) = h(T|X) \cdot \exp(-H(T|X)).$$

Ainsi, les modèles de risques correspondent à la définition d'une fonction de densité pour la variable de durée.

Supposons que L est la vraisemblance d'un échantillon épuré d'observations de la durée comportant des covariables qui changent avec le temps ($X(t)$) et des coefficients de régression (β). La contribution du répondant i à la vraisemblance L compte tenu de β est:

$$L_i = f(T_i|X_i(T), \beta)^{1-c_i} \cdot (1-F(T_i|X_i(T), \beta))^{c_i}$$

où T_i est la date de l'événement ou la date de l'épuration pour le répondant i. La valeur $c_i = 1$ représente l'épuration tandis que $c_i = 0$ correspond à un événement observé. La maximisation de $\log(L) = \Sigma_i \log(L_i)$ (c.-à-d. pour l'échantillon de n répondants) équivaut à la maximisation des risques associés aux événements observés et à la minimisation des risques cumulatifs pour les observations épurées. Après la restructuration, la vraisemblance logarithmique est:

$$\log(L_i) = (1 - c_i)\log(h(T_i|X_i\ (T), \beta)) - \int_0^{T_i} h(t|X_i(t), \beta)\ dt.$$

La régression de Poisson permet de maximiser de façon approximative les résultats de la maximisation de L. L'intervalle $(0, T_i)$ peut être divisé en petits intervalles qui ne se chevauchent pas $((t_{i,j-1}, t_{i,j}]$ (pour la présente étude, les intervalles sont des mois). Les pseudo-observations $e_{i,j}$ indiquent alors si un événement a eu lieu ou non dans chaque intervalle j pour chaque répondant i:

$$e_{i,j} = \begin{cases} 0, \text{ si } t_{i,j} < T_i \\ 0, \text{ si } t_{i,j-1} < T_i < t_{i,j}\ \&\ c_i = 1 \\ 1, \text{ si } t_{i,j-1} < T_i < t_{i,j}\ \&\ c_i = 0. \end{cases}$$

Pour les besoins de l'estimation, l'équation de régression établit un rapport entre les variables indépendantes qui changent avec le temps ($X_{i,j}$) et les espérances mathématiques des pseudo-observations:

$$E(e_{i,j}) = h_{i,j} = \exp(X_{i,j}\ \beta),$$

alors

$$\text{Log}(L) \doteq \Sigma_{i,j} \{ [e_{i,j} \cdot \log(h_{i,j})] - h_{i,j} \},$$

où $h_{i,j}$ est une approximation constante par tranches de la fonction de risque (c.-à-d. dont on sait qu'elle dépend des covariables et de β). La vraisemblance approximative a la même forme qu'une vraisemblance de Poisson (c.-à-d. pour les pseudo- observations $e_{i,j}$), et peut être maximisée à l'aide des moindres carrés pondérés de nouveau de façon itérative (McCullagh et Nelder, 1983) avec un logiciel de régression standard.

4. QUELQUES RÉSULTATS DE LA REGRESSION

Diverses définitions de régression ont été appliquées pour chacun des types d'événements déterminés dans la figure 1. Etant donné qu'au total, 34 équations ont été évaluées, il est impossible de fournir tous les détails s'y rapportant ici (mais ces renseignements sont disponibles sur demande). Les résultats indiqués dans la présente section correspondent à certains aspects des équations les mieux ajustées et portent principalement sur deux points: (i) l'effet de l'entrée sur le marché du travail sur la formation d'unions légitimes et libres et (ii) l'effet de la présence d'enfants sur les risques de séparation. Les équations les mieux ajustées ont été choisies en fonction de la vraisemblance maximisée et à la suite de l'examen des résiduels généralisés (Cox et Oakes, 1984, pp. 88-89).

Une indication du contenu des régressions les mieux ajustées est donnée dans le tableau 2 qui fournit des estimations des taux de transition annuels pour certaines combinaisons de variables indépendantes. Ce tableau donne des taux de base qui facilitent l'interprétation des tableaux subséquents sur les risques relatifs (RR) (c.-à-d. les rapports des risques).

Les taux de base du tableau 1 se comparent assez bien aux estimations appropriées par âge fondées sur la statistique de l'état civil; toutefois, il n'est pas possible de faire d'estimations des statistiques de l'état civil avec des définitions comparables.

4.1 Formation de mariages et d'ULI au moment de l'entrée sur le marché du travail

L'analyse des données de l'EF semble révéler un effet d'indépendance selon lequel l'entrée sur le marché du travail (et l'indépendance financière qui en découle) correspond au taux de mariage le plus élevé qu'une cohorte des périodes prolongées des interruptions du travail (un an et plus). Les questions portant sur l'emploi étaient différentes des autres du fait qu'elles visaient à déterminer le nombre d'années d'activité et la durée en années des arrêts de travail (plutôt que le mois et l'année de l'événement). L'ouvrage de Burch (1975) contient des détails additionnels ainsi qu'un aperçu général du contenu de l'enquête.

Les données de l'EF correspondent à des dates déclarées par chaque répondant pour 0-3 mariages, 0-6 ULI, 0-15 enfants et 0-4 interruptions du travail. En général, ces données étaient suffisantes pour déterminer la situation d'un répondant relativement à son état matrimonial, à la garde d'enfants et à son activité sur une base mensuelle. En vue de l'analyse, le fichier des répondants a été converti en un fichier mensuel dans lequel chaque mois-personne (après le 15^e anniversaire) représentait un enregistrement distinct. A la suite de cette conversion, le fichier de l'EF est passé de 14,004 répondants à plus de 3.9 millions d'enregistrements de mois-personnes.

Le tableau 1 présente les totalisations des chiffres non pondérés des transitions de l'état matrimonial, des répondants et des années-personnes à risque. Ces données et tous les résultats subséquemment utilisés dans le présent rapport ne tiennent pas compte des coefficients de pondération de l'enquête. Selon Hoem (1985, p. 258):

"Si le [plan de sondage] n'apporte pas d'information, on peut alors ne pas tenir compte du plan d'échantillonnage et considérer l'échantillon des événements chronologiques comme autant de voies indépendantes de processus stochastiques avec les caractéristiques probabilistes qu'ils auraient eu sans l'effet du sondage".

(traduction)

Lorsque les coefficients de pondération ont été utilisés pour évaluer les chiffres totaux connus de population à partir des données de l'EF (par ex., les divorces au Canada selon l'année, Burch (1985)), les estimations se sont avérées médiocres. Conséquemment, on a d'abord tenté de faire une analyse des données non pondérées, et les résultats obtenus ont montré que la différence entre les régressions pondérées et les régressions non pondérées n'est pas marquée.

TABLEAU 1
Chiffres non pondérés

Population à risque	Répondants	Années-personnes à risque[3] (mois / 12)	Résultats des transitions	
CEL			ULI	MAR
Jamais travaillé	6545	22863.0	277	656
Déjà travaillé	6795	31167.0	1175	3311
ULI			CEL	MAR
Jamais travaillé	trop peu de cas		.	-
Déjà travaillé	1202	3013.0	303	562
MAR			SEP	
	10456	181560.6	1389	
SEP			DIV	
	1248	4080.4	303	

Il convient de signaler que la structure transitionnelle de la figure·1 a été précisée davantage pour permettre de faire la distinction entre les périodes précédant ou suivant l'entrée sur le marché du travail de chaque répondant.

3. REGRESSION PAR LES RISQUES PROPORTIONNELS

Les modèles de risques représentent des probabilités de transition en temps continu (Cox et Oakes, 1984). Les fonctions de densité et de distribution cumulative (qui dépendent du vecteur de covariables X) de la variable de durée T sont $f(T|X)$ et $F(T|X)$ respectivement. La probabilité de transition conditionnelle (risque) est définie de la façon suivante:

$$h(T|X) \equiv f(T|X)/(1-F(T|X)) = -d \log(1-F(T|X) \quad)/dT$$

(c.-à-d. la limite d'un rapport événement survenu/exposition à mesure que la durée d'exposition approche 0). A partir de cette définition, il s'ensuit que:

$$F(T|X) = 1 - \exp(- \int_0^T h(t|X) \, dt) = 1 - \exp(-H(T|X)),$$

où $H(T|X)$ est la fonction de risque cumulatif et conséquemment:

augmente le risque d'une rupture matrimoniale. Il semble donc essentiel de bien tenir compte, dans l'analyse de la chronologie des événements matrimoniaux, de la possibilité d'effets variant avec le temps et causés par des événements reliés se produisant en parallèle dans la vie des particuliers.

Afin de représenter la dynamique des transitions de l'état matrimonial, chacune des cellules de la matrice de transition correspond à une équation de régression distincte qui tient compte simultanément du vieillissement et des effets des circonstances changeantes. Des équations distinctes mais pas nécessairement indépendantes sont utilisées pour les transitions comportant des risques incompatibles (par ex., de CEL à MAR ou à ULI).

Le diagramme des coefficients de direction montre aussi qu'il existe un besoin en données détaillées sur la situation dans le temps des événements. Par exemple, pour éclaircir la question à savoir si la formation d'une ULI influe sur l'entrée sur le marché du travail ou vice versa, l'ordre de succession de ces événements dans le temps doit être établi. Pour ce faire, les données sur les événements doivent souvent être recueillies mensuellement plutôt qu'annuellement au moyen d'enquêtes approfondies et spécialisées telles que l'enquête sur la famille de 1984. Le manque de données de ce genre a été invoqué comme étant la principale raison pour laquelle notre compréhension de l'évolution du mariage et du divorce au Canada est "rudimentaire" (Balakrishnan et autres, 1987).

1.2 Conclusions d'études antérieures

Dans de nombreuses études du mariage et de la rupture matrimoniale, les covariables examinées étaient soit des caractéristiques individuelles fixes (c.-à-d. ne variant pas dans le temps) comme la race ou l'origine ethnique, soit des caractéristiques non fixes, mais traitées comme telles (par ex., le fait de demeurer en milieu rural ou urbain ou le niveau de scolarité). Cette restriction peut sérieusement réduire la valeur des résultats de certaines études, vu la dynamique illustrée dans la figure 2.

Dans une étude fondée sur les données de recensements récents du Canada, Grenier et autres (1987) ont constaté que la langue maternelle, le lieu de naissance et de résidence, la religion et la scolarité influent de façon significative sur l'âge moyen au mariage. De plus, les effets ont différé entre les recensements de 1971 et de 1981 de même que parmi les cohortes des naissances.

La situation socio-économique des parents a été sérieusement envisagée comme un prédicteur permettant de situer les mariages dans le temps (Hogan, 1978), mais a généralement donné des résultats peu concluants. Il y a cependant un lien étroit entre le niveau de scolarité et le choix du moment du mariage[2]; en effet, plus ce niveau est élevé, plus le mariage est retardé. L'activité et le revenu ont été considérés comme des facteurs ayant une incidence sur les décisions relatives au mariage, bien que la portée de cette incidence diffère selon le sexe (Teachman et autres, 1987). Jusqu'à présent, les résultats laissent entendre que la réussite professionnelle peut avancer le moment du mariage chez les hommes, mais le retarder chez les femmes.

Les statistiques du divorce reçoivent habituellement plus d'attention que celles du mariage. Cette situation est sans aucun doute due au fait que le divorce constitue encore un événement inhabituel (c.-à-d. que jusqu'à maintenant, même les estimations les plus pessimistes indiquent que la plupart des mariages n'aboutissent pas au divorce) alors que le mariage est un fait presque universel (à la longue, 90 à 95% des gens se marient). En outre, la hausse des taux de divorce au Canada après 1968 (réforme de la Loi sur le divorce) peut signifier un important changement social.

Balakrishnan et autres (1987) utilisent les données de l'enquête canadienne sur la fécondité pour déterminer dans quelle mesure l'âge au mariage, la cohorte des mariages, la cohabitation avant le mariage, la situation relative à la fécondité avant le mariage, la pratique religieuse et le lieu de résidence (c.-à-d. la catégorie de taille de la région urbaine) influent sur le risque d'une rupture. Des études de données américaines ont révélé des effets attribuables à la présence d'enfants et à l'emploi ou au revenu (Teachman et autres, 1987). Les conclusions de ces études sont souvent contradictoires. Toutefois, Hannan et Tuma (1978) ont examiné l'incidence de l'augmentation du revenu familial et ont réussi à faire la distinction entre les effets découlant de l'amélioration du bien-être économique d'une famille (c.-à-d. l'augmentation du risque de rupture) et les effets de la réduction de la dépendance économique d'un conjoint sur l'autre (c.-à-d. la diminution du risque de rupture).

2. DONNÈES DE L'ENQUÊTE SUR LA FAMILLE

L'enquête sur la famille de 1984 (EF) fournit pour la première fois des données sur la chronologie d'événements parallèles relatifs à l'état matrimonial, à la fécondité et à l'activité d'un échantillon probabiliste de Canadiens.

L'EF a été menée par Statistique Canada à titre de complément à l'enquête sur la population active. Environ 14,000 Canadiens et Canadiennes de 18 à 64 ans ont été interviewés individuellement et devaient fournir des renseignements chronologiques détaillés sur leurs mariages, leurs unions libres, leurs séparations et leurs divorces (le mois et l'année au cours desquels ces événements avaient eu lieu). En outre, des données ont été recueillies sur les dates du début et de la fin des interruptions du travail pour la garde d'enfants (enfants naturels ou adoptés ou beaux-fils et belles-filles), du 1^{er} emploi du répondant (à plein temps ou à temps partiel, mais d'une durée de six mois ou plus et excluant les emplois occupés par les répondants pendant leurs études) et probabilités. Un changement de situation peut aussi influer sur les probabilités. Par exemple, une personne est plus susceptible de se marier après son entrée sur le marché du travail qu'avant, quel que soit son âge. Dans une étude de la dissolution des unions conjugales chez les femmes suédoises, Hoem (1989, p. 2) souligne les rapports entre les principaux événements de la vie:

FIGURE 1

Matrice de transition

ÉTAT INITIAL	ÉTATS SUBSÉQUENTS				
	CEL	ULI	MAR	SEP	DIV
CEL	X	X	X	-	-
ULI	X	X	X	-	-
MAR	-	-	X	X	-
SEP	-	-	-	X	X
DIV	-	-	-	-	X

"A mesure que sa vie se déroule, la femme modifie sans cesse sa stratégie de vie en fonction d'un système hautement dynamique de ressources, d'expériences, de choix, de restrictions et d'événements fortuits. Sa stratégie régit son comportement et donne lieu à ce qui paraît dans le plan de probabilité comme une relation de causalité réciproque entre son cheminement scolaire et professionnel et ses antécédents familiaux".

Le diagramme des coefficients de direction ci-dessous (figure 2) illustre quelques possibilités relativement à la succession des événements (par ex., une 1re conception avant le mariage) et les voies possibles d'influence (par ex., le retour au travail à la suite d'une séparation). Ce diagramme n'est pas censé représenter des situations typiques. Il existe de nombreux autres ordres de succession et voies d'influence, chacun pouvant exprimer une stratégie de vie différente.

FIGURE 2

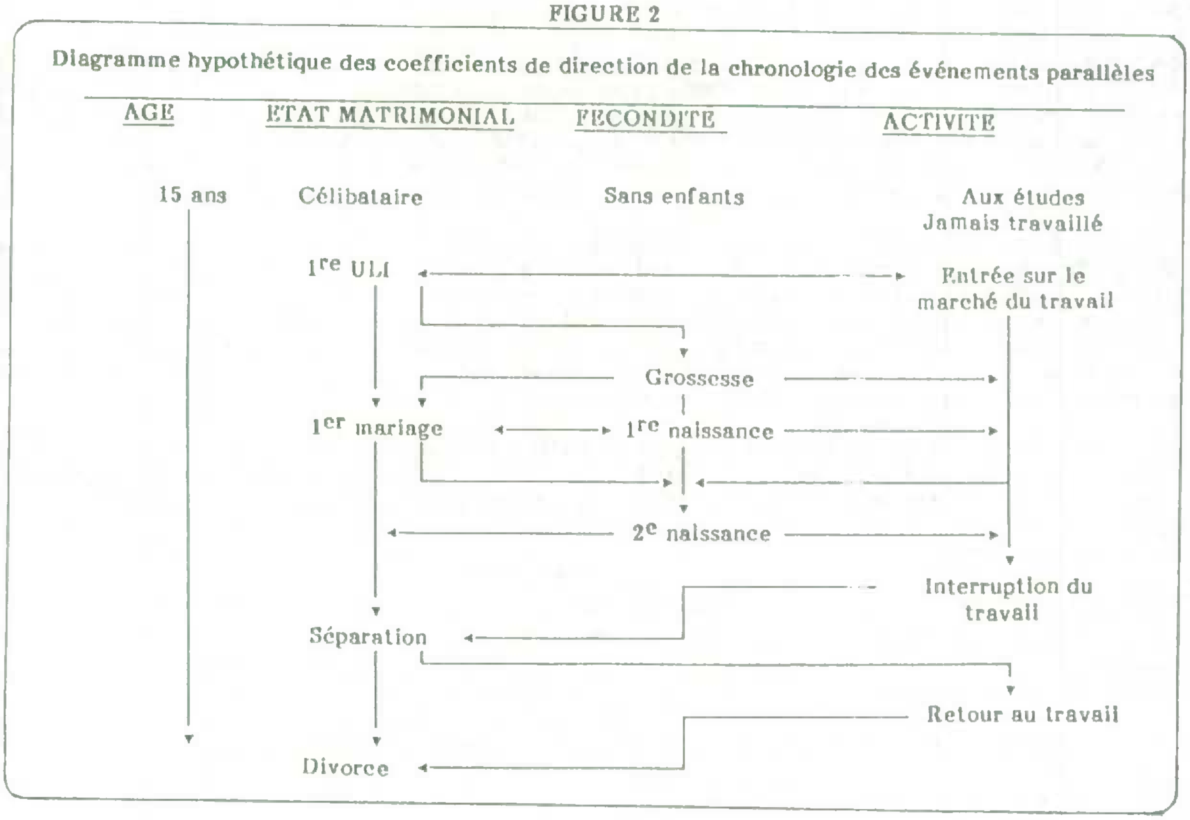


Diagramme hypothétique des coefficients de direction de la chronologie des événements parallèles

De façon générale, le premier mariage est contracté durant la décennie qui marque la fin des études et l'entrée initiale sur le marché du travail, ces événements entraînant souvent en même temps le départ du foyer familial.

Chacun de ces faits représente une étape importante dans le cours de la vie. De même, la situation après le mariage peut changer par suite de la naissance d'enfants ou de l'accumulation d'expériences relatives à l'activité des conjoints. Par exemple, des périodes de chômage peuvent avoir un effet passager ou cumulatif qui

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ANALYSE DE LA CHRONOLOGIE DES MARIAGES ET DES DIVORCES AU CANADA

G. Rowe[1]

RÉSUMÉ

L'enquête sur la famille de 1984 fournit des données qui situent dans le temps les mariages, les naissances, les divorces et les arrêts de travail d'un échantillon probabiliste de Canadiens et de Canadiennes. Les données comprennent les dates des événements s'étant produits entre 1932 et 1984. Les données chronologiques sur les événements individuels reconstitués à partir de ces dates représentent environ 4 millions de mois-personnes d'expérience.

La chronologie des événements matrimoniaux a été analysée à l'aide de la méthode de régression par risques. Les résultats de la régression semblent indiquer qu'il existe un lien entre l'évolution des taux de mariage et de divorce et les régimes d'activité des hommes et des femmes, et que les tendances de la fécondité des mariages influent sur la fréquence des divorces.

Les travaux décrits dans le présent rapport ont été entrepris dans le cadre de l'élaboration d'un modèle de microsimulation démographique (DEMOGEN) à Statistique Canada.

MOTS-CLÉS: Mariage, divorce, chronologie des événements.

1. INTRODUCTION

1.1 Analyse de l'évolution de l'état matrimonial

La plupart des gens se marient tôt ou tard (au moins une fois), mais l'âge au mariage varie considérablement d'une personne à l'autre. Une proportion croissante de mariages se soldent par un divorce, mais les tendances qui se dégagent des ruptures matrimoniales (c.-à-d. les séparations ou les divorces), observées en fonction de la durée du mariage ne sont pas bien comprises (Burch et Madan, 1986). Le présent rapport fait une analyse de la chronologie des événements matrimoniaux qui révèle les liens étroits entre le moment où survient un changement de l'état matrimonial et l'évolution d'autres aspects de la vie des particuliers.

Les descriptions du moment où les événements matrimoniaux se produisent peuvent être données sous forme de probabilités qu'une personne puisse changer d'état matrimonial à intervalles rapprochés. Une telle probabilité sera relativement forte si la durée prévue jusqu'à ce qu'un événement se produise est relativement courte (et, conséquemment, la probabilité sera faible si la durée prévue est longue). Le fait de se fonder sur les probabilités plutôt que sur les durées présente un avantage étant donné qu'un certain événement peut comporter différents types de durées. Par exemple, la durée du mariage et l'âge des enfants (c.-à-d. les intervalles entre les naissances) peuvent influer sur les probabilités de divorce.

La structure des probabilités de transition de l'état matrimonial est représentée par la matrice de transition, à l'intérieur de laquelle certaines combinaisons d'états initiaux et de résultats correspondront à des événements observables tandis que d'autres combinaisons pourront être considérées comme impossibles à réaliser. Dans notre étude, la matrice de transition a la forme présentée dans la figure 1 ci-après pour les états matrimoniaux définis de la façon suivante:

(1) CEL - Jamais marié(e) et ne vivant pas en union libre

(2) ULI - Jamais marié(e) et vivant en union libre

(3) MAR - 1^{er} mariage légitime

(4) SEP - Séparé(e) de façon permanente du conjoint du 1^{er} mariage, mais non divorcé(e)

(5) DIV - Divorce ou annulation du 1^{er} mariage

où les "X" représentent les événements observables et les "-", les événements impossibles.

A priori, les probabilités de transition de l'état matrimonial ne peuvent demeurer constantes avec le temps que dans des circonstances exceptionnelles. L'âge ou la maturité entraînent normalement une modification des

[1] G. Rowe, Division des études sociales et économiques, Statistique Canada, Ottawa, (Ontario) Canada K1A 0T6.

où $B(\theta)$ est une matrice de dimension $(k-1) \times (k-1)$ dont l'élément (i,j) est $q_{ji}(\theta)-q_{ki}(\theta)$, et $C(\theta)$ est une matrice de dimension $(k-1) \times p$ dont la colonne j est égale à $(\partial Q_1'/\partial \theta_j)\pi_1$. On peut s'assurer que les méthodes de notes permettant de résoudre $\partial \log L/\partial \theta = 0$ n'exigent que les premières dérivées de Q et de π_1 par rapport à θ, et, pour cette raison, les algorithmes décrits à la section 4.1 permettent de calculer toutes les quantités nécessaires.

Enfin, nous remarquerons que si les individus sont observés à des moments régulièrement espacés dans le temps, t_0, $t_0 + \tau$, $t_0 + 2\tau$, ..., il est possible d'employer des modèles de chaîne de Markov plus simples (p. ex., Bishop et al. 1975, chap. 7). Dans le cas d'un modèle homogène, ceci permet d'estimer les probabilités de transition $p_{ij}(\tau)$ mais ne fournit pas habituellement des estimations des probabilités de transition plus générales, des intensités ni des distributions des durées de maintien (voir Kalbfleisch et Lawless 1985, section 7). Ceci ne constitue pas nécessairement un inconvénient dans certaines applications.

BIBLIOGRAPHIE

Andersen, P.K. et Borgan, O. (1985). Counting Process Models for Life History Data: A Review (avec discussion). *Scandinavian Journal of Statistics* 12, pp. 97-158.

Andersen, P.K. (1985). Statistical Models for Longitudinal Labor Market Data Based on Counting Processes. Chapitre 6 dans Heckman, J.J. et Singer, B. (ed.) *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.

Bates, D.M. et Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Bishop, Y.M.M., Fienberg, S.E. et Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.

Blumen, J., Kogan, M. et McCarthy, P.J. (1955). *The Industrial Mobility of Labor as a Probability Process*. Cornell Studies of Industrial and Labor Relations, vol. 6. Ithaca, N.Y.: Cornell University Press.

Cinlar, E. (1969). Markov Renewal Theory. *Advances in Applied Probability 1, 123-187*.

Cox, D.R. (1972). Regression Models and Life Tables (Avec discussion). *Journal of the Royal Statistical Society (B)*, 34, 187-220.

Cox, D.R. (1975) Partial Likelihood. *Biometrika*, 62, 269-276.

Cox, D.R. et Miller, H.D. (1985). *The Theory of Stochastic Processes*. London: Methuen (ch. 4).

de Stavola B.L. (1988). Testing Departures From Time Homogeneity in Multistate Markov Processes. *Applied Statistics*, 37, 242-250.

Frydman, Halina (1984). Maximum Likehood Estimation in the Mover- Stayer Model. *Journal of the American Statistical Association*, 79, 632-639.

Hoem, J.M. (1985). Weighting, Misclassification, and Other Issues in the Analysis of Survey Samples of Life Histories, chapitre 5 dans Heckman, J.J. et Singer, B. (ed.). *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge.

Jennrich, Robert I. et Bright, Peter B. (1976). Fitting Systems of Linear Differential Equations Using Computer Generated Exact Derivatives. *Technometrics*, 18, 385-392.

Kalbfleisch, J.D. et Lawless, J.F. (1985). The Analysis of Panel Data Under a Markov Assumption. *Journal of the American Statistical Association*, 80, 863-871.

Kalbfleisch, J.D. et Lawless, J.F. (1988). Likehood Analysis of Multi State Models for Disease Incidence and Mortality. *Statistics in Medicine*, 7, 149-160.

Lawless, J.F. et McLeish, D.L. (1984). The Information in Aggregate Data from Markov Chains. *Biometrika*, 71, 419- 430.

Smith, T.F.M. et Holt, D. (1989). Some Inferential Problems in the Analysis of Surveys Over Time. *Bulletin of the International Statistical Institute*, Vol. 53, 405-424.

Trivellato, U. et Torelli, N. (1989). Analysis of Labor Force Dynamics from Rotating Panel Survey Data. *Bulletin of the International Statistical Institute*, Vol. 53, 425-444.

Tuma, N.B. et Hannan, M.T. (1984). *Social Dynamics; Models and Methods*. New York: Academic Press.

où $u_0 = t_{l-1} < u_1 < \ldots < u_M = t_l$, $\Delta u_i = u_i - u_{i-1}$, la limite étant $M \to \infty$, $\max \Delta u_i \to 0$. Pour obtenir une statistique de caractérisation afin de tester $\gamma=0$, nous avons besoin d'une évaluation de la dérivée de (4.14) par rapport à γ. Il est évident que cette dérivée peut etre exprimée sous la forme :

$$\left.\frac{\partial p_{ij}(t_{l-1},t_l)}{\partial \gamma}\right|_{\gamma=0,\hat{\theta}} = A \int_{t_{l-1}}^{t_l} e^{D(s-t_{l-1})} A^{-1} HAe^{D(t_l-s)} \, ds A^{-1} \tag{4.15}$$

où $D = D(\hat{\theta})$ est la matrice diagonale des valeurs propres pour l'estimation du maximum de vraisemblance dans le modèle homogène ($\gamma=0$) et $Q(\hat{\theta}) = ADA^{-1}$ comme auparavant. On peut encore plus simplifier l'expression (4.15), car la fonction intégrée est une matrice de fonctions exponentielles, et elle peut donc être évaluée simplement.

L'estimation de la variance de la statistique de caractérisation exige l'évaluation de l'hessienne correspondant à (4.15), à $\gamma=0$ et $\hat{\theta}$. Ces calculs sont plus compliqués mais s'effectuent d'une façon semblable à ceux ci-dessus. De Stavola (1988) donne un exemple de cette approche dans un cas simple où les $p_{ij}(s,t)$ peuvent être obtenus algébriquement.

5. QUELQUES AUTRES QUESTIONS

Les études de panel sont principalement utiles pour l'analyse d'histoires personnelles pouvant être modélisées sous la forme de processus de Markov. Autrement, et particulièrement dans les cas où les intensités de transition dépendent fortement du temps passé dans un état, les études de panel ne sont pas particulièrement utiles à moins que les temps d'observation soient proches les uns des autres. Nous remarquons que dans certains types d'études, il peut être possible d'établir rétrospectivement ou au moins d'estimer l'histoire de sujets individuels entre les temps d'observation. On doit alors pondérer l'effet de l'inexactitude possible de certaines informations par rapport aux avantages d'histoires plus complètes.

Plusieurs aspects de l'analyse des données de panel n'ont pas été traités. L'un d'entre eux concerne les enquêtes de panel dans lesquelles, à chaque temps d'observation t_i ($i = 0, 1, \ldots, m$), certains nouveaux individus sont introduits dans l'étude et d'autres en sont éliminés. Par exemple, les enquêtes sur la main- d'oeuvre utilisent souvent des panels dits rotatifs (p. ex., voir Trivelatto et Torelli 1989). Un objectif majeur de telles études consiste souvent à estimer la proportion d'une population occupant divers états à t_0, t_1, ..., t_m, mais au sujet de laquelle il est possible d'obtenir des informations sur les transitions et les temps de maintien dans les états. Un grand nombre d'analyses de telles études reposent sur des méthodes d'enquêtes par sondage (p. ex. Smith et Holt 1989). Il serait avantageux qu'un plus grand nombre d'analyses fondées sur des modèles historiques personnels soient effectuées. Il se présenterait alors des questions de conception intéressantes, concernant par exemple des comparaisons entre l'emploi des données transversales et des données longitudinales (voir Lawless et McLeish 1984).

Les études de panel font normalement intervenir l'observation d'un nombre assez important d'individus à un nombre relativement peu élevé de temps : t_0, t_1, . . . , t_m. Lorsque les processus sont en équilibre, la répartition initiale des états des individus à t_0 peut renfermer des informations considérables. Ceci est déjà incorporé dans les processus de Markov : l'approche est esquissée pour des modèles homogènes.

Soit $Q(\theta)$ la matrice d'intensité de transition $k \times k$ et $\pi = \pi(\theta)$ la répartition à l'équilibre correspondante $k \times 1$; est la solution unique de $Q'\pi = 0$, $\pi_1 + \ldots + \pi_k = 1$. La fonction de vraisemblance fondée sur les distributions $X_l(t_0)$, ..., $X_l(t_m)$ pour les individus $l = 1$, ..., n est :

$$L(\theta) = \prod_{i=1}^{k} \pi_i(\theta)^{n_i(0)} \prod_{ijr} P_{ij}(w_r;\theta)^{n_{ijr}}, \tag{5.1}$$

dans laquelle la notation de la section 4.1 a été utilisée. Pour maximiser (5.1) en résolvant les équations de vraisemblance, nous devons calculer $\partial\pi/\partial\theta'$. Soit Q_1 la matrice $k \times (k-1)$ correspondant au premières $k-1$ colonnes de Q et soit $\pi_1 = (\pi_1, \ldots, \pi_{k-1})'$. Q_1' peut alors être considéré comme une fonction $F(\theta,\pi_1)$ définissant π_1 implicitement en fonction de θ au moyen de $F(\theta,\pi_1) = 0$, de la façon décrite dans Kalbfleisch et Lawless (1985, annexe B). La différenciation implicite de $F(\theta,\pi_1)$ par rapport à θ indique que :

$$\partial\pi_1/\partial\theta' = -B(\theta)^{-1}C(\theta),$$

Les modèles de la forme (4.5) constituent une extension utile des modèles de Markov homogènes, bien qu'ils risquent dans certains cas de représenter une simplification excessive. Par ailleurs, dans le cas des données de panel, une modélisation plus détaillée des effets aléatoires crée habituellement des problèmes d'estimation tellement difficiles qu'il est tout à fait discutable de même les envisager. Soit, par exemple, un modèle à deux états avec des effets aléatoires α_1 et α_2 tel que, étant donné α_1 et α_2, un individu a des intensités de transition $\alpha_1 q_{12}^0$ et $\alpha_2 q_{21}^0$. Même si l'on suppose que α_1 et α_2 sont indépendants, l'estimation est rébarbative. En pratique, α_1 et α_2 sont ne sont pas habituellement indépendants, et l'estimation serait compliquée même pour les histoires personnelles observées continuellement.

4.3 Incorporation du comportement non homogène

i) Les méthodes peuvent être développées pour permettre l'emploi de certains modèles de Markov non homogènes. Si, par exemple :

$$Q(t \mid Z;\theta) = Q(Z;\theta)h(t), \tag{4.12}$$

alors l'emploi de l'échelle de temps opérationnelle $s = \int_0^t h(u)du$ donne naissance à un processus homogène pour $Y_1(s) = X_1(t)$, et les mêmes calculs peuvent alors être effectués pour un $h(t)$ connu. Si $h(t) = h(t;\lambda)$ dépend d'un vecteur de paramètres λ, on peut alors estimer λ en étudiant la vraisemblance de profil de λ, obtenue par maximisation sur les paramètres de régression θ pour chaque λ donné. Le cas correspondant au modèle (4.12), dans lequel h(t) est arbitraire est une question en suspens qui présente un certain intérêt. Il semble probable qu'il soit possible d'élaborer des méthodes pour le traiter.

ii) Une autre approche visant à incorporer la non-homogénéité consiste à laisser la matrice de Markov Q varier à des temps spécifiés. On suppose donc que :

$$Q(t \mid (Z;\theta) = Q_r(Z;\theta) \qquad a_{r-1} \leq t < a_r$$

où $r = 1, \ldots, s$, $a_0 = 0$ et $a_s = \infty$. Les calculs sont simplifiés au maximum si l'on suppose que les changements se produisent à certains des temps d'observation $t_1, \ldots, t_m$, mais ce modèle peut être ajusté de façon assez générale. On peut ainsi incorporer une matrice d'intensités de référence qui varie en fonction du temps; un modèle de la forme (2.5) peut être ajusté.

iii) Des tests d'homogénéité temporelle pourraient être fondés sur n'importe lequel des modèles non homogènes traités ci-dessus. Une autre possibilité découle d'une suggestion de de Stavola (1988), qui a traité le cas spécial suivant. Supposons qu'il n'y ait aucune variable indépendante et examinons un modèle de la forme :

$$Q(t) = Q + H\,\gamma t$$

où $Q = (q_{ij}(\theta))$ comme auparavant et H est une matrice donnée pouvant être non homogène, qui renferme les composantes de $Q(t)$. Par exemple, si l'on pense que les taux de transition de r à s varient avec le temps, on peut poser $H = (h_{ij})_{kxk}$ où $h_{rs}=1$, $h_{rr}=-1$ et $h_{ij}=0$ autrement. On peut alors utiliser un test de notes de $\gamma=0$ pour obtenir une évaluation de l'hypothèse d'homogénéité temporelle par opposition aux cas où le taux diminue ou augmente en fonction du temps.

Si n_{ijl} est le nombre de transitions observées de i à j sur l'intervalle t_{l-1}, t_l, la fonction logarithmique de vraisemblance est :

$$\log L = \sum_{i,j,l} n_{ijl} \log p_{ij}(t_{l-1}, t_l). \tag{4.13}$$

Les probabilités de transition peuvent être exprimées sous la forme d'intégrales de produit :

$$P(t_{l-1}, t_l) = \prod_{u\,\varepsilon(t_{l-1},t_l)} \{I + Qdu + H\,\gamma u du\} \tag{4.14}$$

$$= \lim \prod_{i=1}^{M} \{I + [Q + \gamma H u_i]\,\Delta\,u_i\}$$

Si les $G_i(\alpha)$ sont discrets, la maximisation de la vraisemblance découlant de (4.8) peut être traitée avec les algorithmes présentés à la section 4.1. Une illustration simple mais utile consiste à examiner le modèle mobile-fixe, correspondant au cas spécial où $\alpha = 0,1$ seulement. Blumen et al. (1955) ont introduit la version temporelle discrète et Frydman (1984) a traité l'estimation de la vraisemblance maximale dans le modèle discret. Nous définissons $s_i = \Pr\{\alpha_1 = 0 \mid X_1(t_0) = i\} = 1-\Pr\{\alpha_1=1 \mid X_1(t_0) = i\}$ et nous examinons l'estimation combinée de $s = (s_1, \ldots, s_k)$ et de θ. La fonction de vraisemblance est le produit pour $1 = 1, \ldots, N$ des termes (4.8). En notant que $p_{ij}(0;\theta) = \delta_{ij} = I(i=j)$ et en définissant $n_{ijr} = \#\{1:X_1(r-1) = i, X_1(r) = j\}$, $n_i^\star = \#\{1:X_1(r) = i \text{ pour } r = 0,1, \ldots, m\}$ et $N_{ir} = \#\{1:X_1(r) = i\}$, nous pouvons exprimer la vraisemblance sous la forme

$$L(\theta,s) = \prod_{i,j=1}^{k} \prod_{r=1}^{m} p_{ijr}^{n_{ijr}-\delta_{ij}n_i^\star} \prod_{i=1}^{k} [s_i+(1-s_i)H_i]^{n_i^\star} (1-s_i)N_{i0}-n_i^\star \tag{4.9}$$

où p_{ijr} désigne $p_{ij}(w_r;\theta)$ et $H_i = \prod_{r=1}^{m} p_{ii}(w_r;\theta)$.

On peut montrer que les premières dérivées de log L sont égales à

$$\frac{\partial \log L}{\partial\theta} = \sum_{ijr} n_{ijr}\, p_{ijr}^{-1}\, p'_{ijr} - \sum_i \frac{n_i^\star\, s_i}{s_i + (1-s_i)\, H_i} (\partial \log H_i/\partial\theta) \tag{4.10}$$

$$\frac{\partial \log L}{\partial s_i} = \frac{n_i^\star(1-H_i)}{s_i + (1-s_i)\, H_i} - \frac{N_{i0}-n_i^\star}{1-s_i}, \tag{4.11}$$

où $p'_{ijr} = \partial p_{ijr}/\partial\theta$, et nous remarquons que $\partial \log H_i/\partial\theta = \sum_{r=1}^{m} p_{iir}^{-1}\, p'_{iir}$. Toutes les quantités dans (4.9) et (4.10) peuvent être obtenues avec les algorithmes présentés dans Kalbfleisch et Lawless (1985), qui ont déjà été mentionnés à la section 4.1.

Les composantes de la matrice d'information de Fisher peuvent être calculés directement :

$$E(-\frac{\partial^2 \log L}{\partial s_i^2}) = \frac{N_{i0}(1-H_i)}{(1-s_i)(s_i+(1-s_i)H_i)}$$

$$E(-\frac{\partial^2 \log L}{\partial s_i \partial\theta}) = \frac{N_{i0}H_i}{s_i+(1-s_i)H_i} (\partial \log\, H_i/\partial\theta)$$

$$E(-\frac{\partial^2 \log L}{\partial\theta\partial\theta'} = \sum_{i,j,1} (E(N_{i,1-1}) - N_{i0}s_i)\, p_{ij1}^{-1}\, p'_{ij1}(p'_{ij1})^t$$

$$- \sum_i \frac{N_{i0}s_i(1-s_i)H_i}{s_i+(1-s_i)H_i} (\partial \log H_i/\partial\theta)(\partial \log H_i/\partial\theta)^t.$$

On en obtient facilement une estimation en remplaçant $n_{i,1-1}$ par $E(N_{i,1-1})$.

Comme dans le cas homogène de la section 4.1, l'algorithme de notes de Fisher constitue une méthode simple d'ajustement des données, dans laquelle on a seulement besoin des premières dérivées des $p_{ij}(w_r;\theta)$ par rapport à θ. On remarquera que $\partial \log L/\partial s_i = 0$ donne les s_i en fonction de θ, ce qui permet d'effectuer certaines simplifications.

Des modèles mobiles-fixes plus généraux peuvent aussi être ajustés. Ils permettent par exemple d'établir les dépendances de s_i par rapport aux variables indépendantes, ou encore des modèles de régression des intensités de transition q_{ij}.

où $D = \text{diag}\ (d_1, \ldots, d_k)$. Cela permet de calculer simplement les composantes $p_{ij}(t \mid Z,\theta)$ de la vraisemblance et le vecteur de notes $\partial \log L/\partial\theta$. Les premières dérivées du vecteur de notes,

$$\frac{\partial}{\partial\theta}\, p_{ij}\ (t \mid Z;\theta) \tag{4.4}$$

peuvent également être obtenus à l'aide d'un algorithme proposé par Jennrich et Bright (1976). Bates et Watts (1988, annexe A) traitent des méthodes de calcul dans lesquelles les valeurs propres ne sont pas distinctes. Kalbfleisch et Lawless (1985) ont montré qu'une variante de l'algorithme de notes permettant de trouver les estimations du maximum de vraisemblance, exige seulement le calcul des premières dérivées (4.4). À la convergence, on dispose d'un estimateur de variance simple pour $\hat{\theta}$.

On remarquera que ces méthodes exigent que l'on calcule séparément chacune des valeurs distinctes des variables indépendantes. Donc, si le nombre d'ensembles de variables indépendantes distinctes de l'échantillon est grand, ces méthodes exigent beaucoup de calculs. Toutefois, les données de panel ne conviennent souvent pas à la modélisation très détaillée des effets des variables indépendantes, et, dans bien des cas, le volume des calculs n'est pas excessif.

4.2 Incorporation de l'hétérogénéité non observable

Même après avoir modélisé la dépendance des intensités de transition à l'égard des variables explicatives, nous risquons de constater qu'un modèle homogène de Markov est inadéquat. Si les modèles de Markov pour les individus sont considérés raisonnables, une approche consiste à modéliser l'hétérogénéité non observable des intensités de transition individuelles au moyen d'effets aléatoires. Nous indiquons comment certains modèles simples mais utiles peuvent être ajustés.

Un modèle de Markov complet des intensités de transition individuelles utiliserait des matrices d'intentités de transition $Q(Z_l;\theta \mid \alpha_l)$ où α_l est un vecteur non observable d'effets aléatoires associés à l'individu l. Dans certaines situations, des types spéciaux d'effets aléatoires peuvent se dégager naturellement. Considérons ici seulement la famille simple mais utile de modèles pour lesquels

$$Q(\theta \mid \alpha_l) = \alpha_l Q^0(\theta), \tag{4.5}$$

où $Q^0(\theta)$ est une matrice d'intensité de référence et les α_l sont des variables aléatoires indépendantes caractérisées par une fonction de distribution $G(\cdot)$. Pour simplifier notre exposé, nous avons supposé qu'il n'y avait aucune variable indépendante fixe Z_l.

Pour exprimer la probabilité inconditionnelle

$$\Pr\{X_l(t_0),\ldots,X_l(t_m)\} = \int_0^\infty \Pr\{X_l(t_1),\ldots,X_l(t_m) \mid X_l(t_0),\alpha\}\ \Pr\{X_l(t_0) \mid \alpha\} dG(\alpha) \tag{4.6}$$

nous devons spécifier la répartition combinée de $X_l(t_0)$ et α_l. En définissant $\eta_i = \Pr\{X_l(t_0) = i\}$ et $G_i(\alpha)$ comme la fonction de répartition conditionnelle de α_l étant donné $X_l(t_0) = i$, nous pouvons exprimer (4.6) sous la forme

$$\eta_i \int_0^\infty \Pr\{X_l(t_1),\ \ldots,\ X_l(t_m) \mid X_l(t_0) = i,\ \alpha\}\ dG_i(\alpha) \tag{4.7}$$

lorsque $X_l(t_0) = i$. Dans ce qui suit, nous n'avons pas tenu compte des η_i et de l'estimation θ, pas plus que des paramètres des $G_i(\alpha)$ dans la vraisemblance résultant du second terme de (4.7). Le modèle (4.5) suppose que

$$P(t;\theta \mid \alpha) = P^0(\alpha t;\theta),$$

où $P^0(t;\theta) = \exp\{Q^0(\theta)t\}$. La contribution à la vraisemblance de l'individu l est donc

$$L_l = \int_0^\infty \{\prod_{r=1}^{m} p^0_{i_{r-1}i_r}\ (\alpha w_r)\}\ dG_i(\alpha), \tag{4.8}$$

où $i_{r-1} = x_l(t_{r-1})$, $i_r = x_l(t_r)$ et $w_r = t_r - t_{r-1}$ $(r=1,\ \ldots,m)$. On remarquera toutefois que la modélisation des $G_i(\alpha)$ n'est pas toujours claire dans de nombreuses situations.

indépendantes qui varient avec t (p. ex., conditions économiques) et de l'histoire personnelle de l'individu, soit par exemple la durée de maintien dans l'état couramment occupé. Ces variables peuvent facilement être étudiées dans le modèle (3.1). Le but de cet article n'est pas d'étudier des méthodes d'observation continue, mais plutôt de préciser plusieurs secteurs dans lesquels des études supplémentaires seraient utiles :

i) Dans de nombreuses applications, les histoires personnelles sont partiellement déterminées rétrospectivement. Par exemple, des individus peuvent être échantillonnés à un certain temps et leur histoire personnelle récente reconstruite, ce qui serait par exemple le cas de la détermination de la durée de la période de chômage courante ou de l'utilisation récente d'établissements hospitaliers. En plus des problèmes de précision des données, il est essentiel de tenir compte du biais de sélection en utilisant des vraisemblances appropriées (p. ex., voir Hoem, 1985, Kalbfleisch et Lawless, 1988). En outre, il existe des problèmes de conception et d'analyse.

ii) On peut incorporer l'hétérogénéité non observable dans des modèles tels que (3.1), en multipliant par exemple les intensités de transition par des variables aléatoires α_{ij}. Toutefois, même pour les modèles très simples, cela risque d'être plutôt compliqué; Andersen (1985) propose une approche faisant appel à l'analyse fondée sur la vraisemblance partielle. Un autre problème est l'impossibilité de distinguer entre l'hétérogénéité non observable et certains types de dépendances temporelles dans une population homogène; d'autres éclaircissements sur cette question seraient précieux.

iii) Des études plus poussées et l'application de modèles incorporant une dépendance du temps civil et de la durée de maintien dans l'état courant seraient précieuses. Les expériences d'utilisation de tels modèles dans des processus sociaux et économiques sont à l'heure actuelle limitées.

4. DONNÉES DE PANEL

4.1 Estimation des processus homogènes de Markov

La gamme étendue des techniques d'analyse des données continues contraste avec la gamme limitée des techniques d'analyse des données de panel. Dans cette section, nous avons traité des méthodes fondées sur les modèles homogènes de Markov ainsi que sur certaines de leurs extensions simples.

Soit Z_l un vecteur de variables indépendantes $(Z_{l1} = 1, \ldots, Z_{lp})$ associé à l'individu l. Le processus $X_l(t)$ est considéré comme un processus homogène de Markov dont les intensités de transition sont

$$q_{ij}(Z_l) = \exp(Z_l'\beta_{ij}), \; i \neq j \tag{4.1}$$

où $\beta_{ij}' = (\beta_{ij1}, \ldots, \beta_{ijp})$ est un vecteur de coefficients de régression. On remarquera que exp (β_{ij1}) est l'intensité de transition de référence lorsque $Z_{l2} = \ldots = Z_{lp} = 0$. Nous supposons que les vecteurs β_{ij} sont des fonctions du vecteur de paramètres $\theta = (\theta_1, \ldots, \theta_q)$. Dans la plupart des applications, seulement quelques composantes de β_{ij} ne sont pas nulles, et, généralement, θ est de dimensions modérées.

Soit la matrice de probabilités de transition définie sur un intervalle de longueur t

$$P(t \mid Z;\theta) = \exp(Q(Z;\theta)t) = \sum_{j=0}^{\infty} \{Q(Z,\theta)t\}^j/j! \tag{4.2}$$

Enfin, soit F_{ijl} l'ensemble des individus observés qui sont passés de i à j dans l'intervalle (t_{l-1}, t_l), $l = 1,$ ..., m, i,j = 1, . . ., k. La vraisemblance basée sur ces données est :

$$L(\theta) = \prod_{l=1}^{m} \prod_{i=1}^{k} \prod_{j=1}^{k} \prod_{r\varepsilon F_{ijl}} p_{ij}(t_l - t_{l-1} \mid Z_r;\theta). \tag{4.3}$$

Pour obtenir une estimation de la vraisemblance maximale de θ à partir de (4.3), nous avons utilisé l'algorithme présenté dans Kalbfleisch et Lawless (1985). Pour une valeur donnée de Z et une valeur spécifiée de θ, supposons que $Q(Z;\theta)$ a des valeurs propres distinctes : $d_1 = 0$, $d_j(Z;\theta)$, j=2, ..., k. Donc

$$Q(Z;\theta) = A(Z;\theta)\; D(Z;\theta)\; A(Z;\theta)^{-1}$$

où les colonnes de A sont les valeurs propres exactes de $Q(Z;\theta)$ et

$$P(t \mid Z;\theta) = A(Z,\theta)e^{D(Z,\theta)t}\, A(Z;\theta)^{-1}$$

temps écoulé depuis le début du processus. Le cas spécial d'un processus de Markov homogène est caractérisé par :

$$q_{ij}(t) = q_{ij} \tag{2.2}$$

indépendant de t, et il s'avèrera un modèle particulièrement utile pour les données de panel. Il est commode de poser $q_{ii}(t) = -\sum_{j\neq i} q_{ij}(t)$, $i = 1, \ldots, k$ et $Q(t) = (q_{ij}(t))_{k\times k}$ dans le cas non homogène ou $Q = (q_{ij})_{k\times k}$ dans le cas homogène. Pour les processus homogènes, on peut démontrer que :

$$P(t) = e^{Qt} = I + Qt + Q^2 t^2/2! + \ldots \tag{2.3}$$

où $P(t) = (p_{ij}(t))_{k\times k}$ et $p_{ij}(t) = P\{X(t) = j \mid X(0) = i\}$. Voir, par exemple, Cox et Miller (1965).

Dans le processus semi-markovien, on a supposé que les intensités de transition dépendaient uniquement du temps écoulé dans l'état courant. Donc :

$$\lambda_j(t;\, J_r,\, T_r,\, r = 0, \ldots, M(t)) = \pi_{ij}(x) \tag{2.4}$$

où $J_{M(t)} = i$ et $x = t - T_{M(t)}$. De tels processus sont par exemple décrits dans Cinlar (1969).

On peut également étudier des modèles plus généraux. Par exemple, les intensités de transition peuvent être fonction des durées des états x et du temps civil t. De tels processus sont dits semi- markoviens non homogènes et ils fournissent une classe très souple de modèles. Nous avons étudié par la suite un modèle fixe-mobile, dans lequel nous supposons que chaque individu dans l'état i au temps 0 a une chance s_i de rester dans cet état pendant la durée complète. Avec la probabilité complémentaire $1 - s_i$, $X(t)$ est un processus de Markov dont les intensités sont $q_{ij}(t)$.

Il existe des extensions aux modèles de régression. Si Z est un vecteur de variables indépendantes, il est naturel de considérer les modèles de Markov ayant des intensités

$$q_{ij}(t;Z) = q^0_{ij}(t)\ \exp(Z'\beta_{ij}) \quad i \neq j \tag{2.5}$$

où $q^0_{ij}(t)$ est une fonction d'intensité de référence qui est appliquée lorsque $Z = 0$, et β_{ij} est un vecteur de paramètres de régression. (Il est possible de remplacer $\exp(Z'\beta_{ij})$ dans (2.5) par une fonction de risque relative $r(Z'\beta_{ij})$, mais nous utiliserons dans tous les cas la fonction de risque relative exponentielle.) On obtient les modèles de régression semi-markoviens d'une façon similaire. Plus généralement, ainsi que nous l'avons exposé à la section 3, il est possible de permettre aux variables indépendantes de varier en fonction du temps. Donc, pour les modèles de Markov, nous pouvons envisager une variable indépendante Z_t au temps t, qui peut dépendre du processus jusqu'au temps t. Donc, Z_t peut renfermer des variables indépendantes mesurées, des produits de variables indépendantes et de temps, des informations sur les états précédents occupés, ou encore le temps $x = t - T_{M(t)}$ dans l'état présent.

Avec les données de panel, il est habituellement difficile d'utiliser autre chose que les processus de Markov avec des variables indépendantes fixes Z. Toutefois, avec les histoires personnelles observées continuellement, on peut travailler assez facilement avec une gamme étendue de modèles. Nous allons traiter quelques-unes des méthodes dans la section suivante.

3. MÉTHODES POUR LES HISTOIRES PERSONNELLES CONTINUES

Si les individus sont suivis prospectivement (dans le temps) et que leurs histoires personnelles continues soient observées, il est possible d'obtenir facilement les fonctions de vraisemblance, et les méthodes d'inférence sont simples. En particulier, les modèles de Markov dans (2.5) peuvent être généralisés en des modèles à intensité multiplicative, dans lesquels les intensités ont la forme

$$q_{ij}(t;Z_t) = q^0_{ij}\exp(Z_t'\beta) \tag{3.1}$$

où $q^0_{ij}(t)$ est une intensité de référence et Z_t un vecteur de variables indépendantes pouvant dépendre du temps. Avec ce modèle, il est possible d'utiliser des analyses fondées sur la vraisemblance partielle, décrites par Cox (1972, 1975). Andersen et Borgan (1985) ont étudié cette question pour les processus historiques personnels, et Andersen (1985) a traité certaines applications économiques. Ces méthodes semblent avoir un vaste domaine d'application dans les études socioéconomiques, en particulier lorsque les variables indépendantes dépendent du temps. Par exemple, dans les études de chômage et d'emploi, les intensités de transition dépendent normalement de variables indépendantes fixes associées à l'individu, du temps civil t, de variables

Recueil du symposium de Statistique Canada
sur l'analyse des données chronologiques
octobre 1989

QUELQUES MÉTHODES STATISTIQUES D'ANALYSE DE DONNÉES HISTORIQUES PERSONNELLES DE PANEL

J.D. Kalbfleisch et J.F. Lawless[1]

1. INTRODUCTION

Dans des disciplines telles que la démographie, l'economie, la médecine et la sociologie, il est courant d'étudier des processus historiques personnels à états multiples d'individus d'une population donnée (p. ex., Hoem 1985, Kalbfleisch et Lawless 1988, Tuma et Hannan 1984). Supposons par exemple que N individus parcourent indépendamment k états $\{1,2, \ldots, k\}$ sur une période de temps donnée. Les états peuvent représenter, entre autres, des maladies, des catégories d'occupations, des situations familiales ou des indicateurs socio-économiques. Soit $X_l(t)$ l'état occupé par l'individu l au temps t et Z_l un vecteur de variables indépendantes observées sur l'individu l. Étant donné $Z_1, \ldots, Z_N$, les processus $\{X_l(t) : 0<t<\infty\}$, $l = 1, \ldots, N$ sont supposés être indépendants.

Parfois, l'histoire complète de la vie d'individus est observée sur certains intervalles de temps. À la section 3, nous faisons brièvement le point sur cette situation. Toutefois, l'objet principal de cet article est d'étudier les situations dans lesquelles l'individu l est observé à un ensemble préétabli de points dans le temps, soit $t_{l0}, \ldots, t_{lm_l}$. Les états $X_l(t_{lj})$ correspondant à ces points sont alors observés. Par ailleurs, on ne dispose d'aucune information sur la trajectoire entre les temps d'observation successifs. Les données de ce type sont habituellement appelées données de panel. Pour faciliter notre exposé, nous avons supposé que $t_{lj}=t_j$, $j=0, \ldots, m$ où $m_l = m$, $l=1, \ldots, N$ et $t_0 = 0$. Toutefois, les méthodes présentées peuvent facilement être généralisées pour correspondre au cas dans lequel les temps d'observation diffèrent d'un individu à l'autre. La modélisation des processus $X_l(t)$ vise à décrire adéquatement la variation observée dans les données et présente donc un intérêt tout à fait particulier. La comparaison des processus pour des individus de groupes différents et l'évaluation des variables indépendantes sont souvent importantes.

Dans la section 2, nous avons traité les processus markoviens et semi-markoviens, ceux-ci étant largement utilisés en tant que modèles des processus historiques personnels. Dans la section 3, nous avons examiné les méthodes existantes correspondant aux situations dans lesquelles les processus $X_l(t)$ sont observés continuellement dans le temps et nous avons abordé certains problèmes supplémentaires. En fait, il existe diverses techniques fondées sur un vaste choix de modèles. Dans la section 4, nous avons traité assez longuement les méthodes s'appliquant aux données de panel. Dans ces situations les méthodes et les modèles commodes sont alors bien plus rares. Notre objectif est de décrire certaines méthodes d'analyse pouvant être mises en pratique; en particulier, nous avons présenté des procédures pour les modèles de Markov homogènes dans le temps ainsi que des extensions simples de ces procédures, y compris l'incorporation de variables indépendantes fixes, de l'hétérogénéité non observable et de la dépendance temporelle. Nous avons conclu à la section 5 en présentant certaines remarques sur l'utilité des données de panel ainsi que sur des problèmes connexes tels que les enquêtes de panel.

2. QUELQUES MODÈLES POUR LES PROCESSUS HISTORIQUES PERSONNELS

Étant donné un processus chronologique continu $\{X(t): 0<t<\infty\}$ défini sur l'espace d'états $S = \{1,2, \ldots, k\}$. Pour des raisons de commodité, nous avons supposé que le processus entre dans l'état J_0 au temps $T_0 = 0$. Soit $M(t)$ le nombre de transitions dans $(0,t)$, T_r le temps de la transition r et J_r l'état occupé immédiatement après la transition r. La modélisation peut être réalisée grâce à l'emploi d'intensités de transition instantanées:

$$\lambda_j(t;J_r,T_r,r = 0, \ldots, M(t)) = \lim_{\Delta t\to 0} P\, \{T_{M(t)+1}\varepsilon(t,t + \Delta t),\ J_{M(t)+1}=j \mid J_r,T_r,r = 0, \ldots, M(t)\}/\Delta t.$$

Deux cas spéciaux présentent un intérêt particulier. Les processus de Markov, pour lesquels

$$\lambda_j(t;\ J_r,T_r,\ r = 0, \ldots, M(t)) = q_{ij}(t) \tag{2.1}$$

où $J_{M(t)} = i \neq j$, spécifient que les intensités de transition dépendent seulement du temps civil, ou encore du

[1] J.D. Kalbfleisch et J.F. Lawless, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, (Ontario), Canada N2L 3G1.

SECTION 6

DÉMOGRAPHIE

BIBLIOGRAPHIE

Balfe, D.L., W.J. Steinberg, H.G.V. Kustner. 1988, "Comparison of the decline in the ischaemic heart disease mortality rate in the RSA with that in other Western Countries". South African Medical Journal, 74; p. 551-553.

Base canadienne de données sur la mortalité, statistique de l'état civil. Centre canadien d'information sur la santé. Statistique Canada.

Blackburn, Henry. 1989, "Trends and determinants of CHD mortality: Changes in risk factors and their effects". International Journal of Epidemiology, 18-S1; p. 210-215.

Castelli, William. 1989, "Determinants of CHD mortality: Genetic: socioeconomic, lifestyle and risk factor influences: An overview". International Journal of Epidemiology, 18-S1; p. 180-182.

Enquête Santé Canada. 1981, La Santé des Canadiens: Rapport de l'Enquête Santé Canada (no. 82-538 au catalogue). Hull Québec. Ministre des Approvisionnement et Services, Canada.

Epstein, Frederick. 1989, "The relationship of lifestyle to international trends in CHD". International Journal of Epidemiology, 18-S1; p. 203-209.

---- 1989, "Trends and determinants of coronary heart disease mortality: International comparisons". International Journal of Epidemiology, 18-S1.

Hughes, L.O., J.H. Cruickshank, J. Wright, E.B. Raftery. 1989, "Disturbances in insulin in British Asian and white men surviving myocardial infarction". British Medical Journal, 299; p. 537-541.

Hughes, L.O., V. Raval, E.B. Raftery. 1989, "First myocardial infarctions in Asians and white men". British Medical Journal, 298; p. 1345-1350.

Marmot, Michael. 1989, "Socioeconomic determinants of CHD mortality". International Journal of Epidemiology, 18-S1; p. 196-202.

McKeigue, P.M., G.J. Miller, M.G. Marmot. 1989, "Coronary heart disease in South Asians overseas: A review". Journal of Clinical Epidemiology, 42-7; p. 597-609.

Ministère de la Santé nationale et du Bien-être social. 1973, Enquête Nutrition Canada (no. H58-36 au catalogue). Information Canada.

Nair, C., H. Colburn, D. McLean et A. Petrasovits. 1989, "Maladies cardio-vasculaires au Canada". Rapports sur la santé, vol. 1, no. 1.

Organisation mondiale de la santé. 1988, "Noncommunicable Diseases: A global problem". Rapport trimestriel de statistiques sanitaires mondiales, 41.

Recensements de la population du Canada (1971 et 1986). Statistique Canada

Salonen, Jukka T. 1989, "Non-insulin dependent diabetes and ischaemic heart disease". British Medical Journal, 298-1050.

Santé et Bien-être social Canada. 1989, Enquête sur la tension artérielle des Canadiens (no. H39-143 au catalogue). Ministre des Approvisionnement et Services, Ottawa.

Statistique Canada. 1985, La population active, juin 1985 (no. 71-001 au catalogue). Ministre des Approvisionnements et Services, Ottawa.

Statistique d'immigration. 1988, Statistiques sur l'immigration (trimestrielles). Emploi et Immigration Canada.

Thom, Thomas, J. 1988, "International mortality from heart disease: Rates and trends". International Journal of Epidemiology, 18-S1; p. 20-28.

5. ASIE

6. CHINE
7. JAPON*
8. ASIE DU SUD
Inde
Pakistan

12. EUROPE DE L'OUEST
Autriche
France
Grèce
Italie
Espagne
Royaume Uni
Belgique
Allemagne
Irlande
Portugal
Suisse

* L'élimination est due aux chiffres peu élevés.

Tableau 2: Disponibilité des données

INTERVALLE	NOMBRE DE DÉCÈS	CHIFFRES DE POPULATION
1964-1968	Disponible	Recensement abrégé / il manque le lieu de naissance
1969-1973	Disponible	Disponible
1974-1978	il manque Ontario, tous les ans C.-B., 3 ans Manitoba, 1 an	Recensement abrégé / il manque le lieu de naissance
1979-1983	il manque; Ontario, 1 an Manitoba, 4 ans Alberta, 2 ans	Disponibles
1984-1988**	il manque; Alberta, 1 an	Recensement abrégé / lieu de naissance disponible

* selon le lieu de naissance, l'âge et le sexe
** il manque tous les chiffres (décès et population) pour 1988

Tableau 3: Ordre des taux de mortalité causée par MCV normalisés selon l'âge - hommes

IMMIGRANTS AU CANADA	PAYS DE NAISSANCE
Scandinavie	Europe de l'Est Scandinavie
Amérique du Nord Europe de l'Est Europe de l'Ouest	Europe de l'Ouest Amérique du Nord
Asie du Sud Chine Amérique du Sud	Amérique du Sud Hong Kong Sri Lanka

Tableau 4: Ordre des taux de mortalité causée par MCV normalisés selon l'âge - femmes

IMMIGRANTES AU CANADA	PAYS DE NAISSANCE
Scandinavie Europe de l'Ouest Europe de l'Est	Europe de l'Est Europe de l'Ouest Scandinavie
Amérique du Nord	Amérique du Nord
Asie du Sud Chine Amérique du Sud	Amérique du Sud Hong Kong Sri Lanka

Hughes *et al.*, 1989a; Hughes *et al.*, 1989b), mais on ignore s'il est de nature causale ou génétique. Les résultats obtenus par McKeigue *et al.*, 1989 et divers autres (Hughes *et al.*, 1989a; Hughes *et al.*, 1989b) ne sont pas compatibles avec ceux de la présente étude des taux de mortalité chez les Canadiens d'origine sud-asiatique de la première génération.

Notre étude montre que, pour les sud-asiatiques, les TMNA pour les MCV ainsi que toutes les autres causes de décès ont été constamment plus bas que la plupart des autres groupes d'immigrants canadiens étudiés. Cependant, elle montre aussi que la proportion de Sud-asiatiques qui meurent de MCV est la plus élevée de tous les groupes pour les années 1984-1988. Ceci nous indique que les Sud-asiatiques sont plus susceptibles de mourir de MCV que de toute autre cause de décès. Les données sur les décès utilisées pour notre étude nous montre que les décès chez les Sud-asiatiques se produisent proportionnellement plus souvent à un âge moins élevé que le reste des Canadiens. Cet écart peut s'expliquer du fait que le processus d'immigration offre de plus grandes chances d'admission aux personnes en meilleure santé et (ou) du fait que la qualité des services de santé disponibles au Canada permet davantage de prolonger la vie des personnes atteintes de MCV.

5. CONCLUSION

Les taux de mortalité causés par MCV varient chez les divers groupes ethniques de la première génération au Canada. Environ un Canadien sur cinq est un immigrant de la première génération, d'où l'importance de considérer l'origine ethnique dans l'interprétation des statistiques de la santé et dans la planification des services de santé au Canada. Le mode de vie joue un rôle essentiel lorsqu'il s'agit d'établir le degré d'exposition aux facteurs de risque et le niveau de santé qui en résulte. La plupart des immigrants tendent à conserver pour le meilleur ou pour le pire, dans leur pays d'adoption, leurs habitudes culturelles, comme le régime alimentaire et la consommation de tabac. Des changements au mode de vie peuvent s'avérer nécessaires afin de réduire les facteurs prédisposant aux MCV et à d'autres affections.

Aux fins de l'analyse des données sur la santé dans une société multiculturelle et multi-ethnique comme celle du Canada, il convient de rassembler des renseignements sur les groupes ethniques dans tous genres de collectes de données sur la santé et l'état civil. L'une des principales difficultés rencontrées au cours de la présente étude a consisté dans de grandes lacunes d'information sur l'origine ethnique. En particulier, l'absence de données a empêché l'examen de la population autochtone du Canada.

L'étude des variations des taux de mortalité d'un pays à l'autre aide depuis longtemps à l'identification des facteurs étiologiques. Une récente publication (Epstein, 1989) traite des tendances et des éléments déterminants de la mortalité causée par MCV à l'échelle internationale. Les facteurs de risques rattachés aux traits génétiques, aux caractéristiques socio-économiques et au mode de vie ont tous une influence sur les variations internationales (Castelli, 1989). Il se peut que les influences génétiques et la variabilité des facteurs prédisposant aux MCV soient très complexes et très difficiles à éclaircir. Il existe un rapport entre les variables socio-économiques et les MCV, les taux étant plus élevés dans les pays d'abondance, quoique dans ces pays, de faibles taux de MCV soient associés au groupe des riches. Si l'on considère l'évolution des classes sociales, une étude de Marmot, 1989 montre qu'en Grande-Bretagne, la progression dans l'échelle sociale peut bénificier à certains groupes d'immigrants tout en nuisant à d'autres.

En ce qui a trait aux facteurs de risque, il est démontré que si la consommation de matières grasses augmente, elle prédispose davantage aux MCV; que si elle demeure inchangée, elle n'occasionne pas de variation; et que si elle est réduite, elle amène une diminution. Cette association n'est pas confondue par la consommation de tabac ou d'alcool (Epstein, 1989). Aux États-Unis, les campagnes de lutte contre l'hypertension se sont traduites par une baisse de 54% du taux de mortalité due aux maladies du coeur (Blackburn, 1989). Les changements de consommation de tabac n'aident pas à expliquer la variation des taux de regression, mais ils y contribuent selon toute probabilité. La situation est très complexe et, pour l'expliquer, on peut trouver utile d'étudier l'évolution des groupes d'immigrants.

Tableau 1: Groupe de Pays

1. **AMÉRIQUE DU NORD**
   Canada
   États-Unis

2. **AMÉRIQUE DU SUD**
   Brésil
   Chili
   Mexique

3. **AFRIQUE**
   Continent

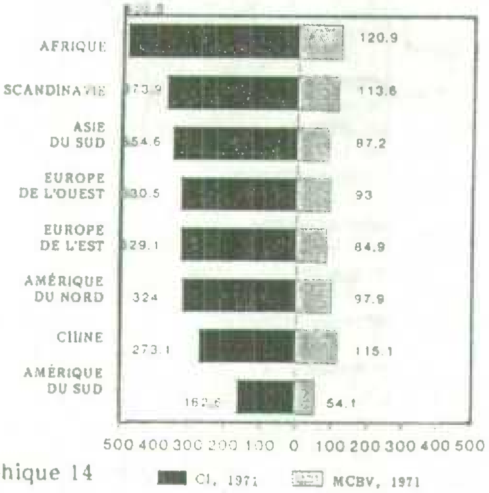
4. **AUSTRALIE***
   Continent

9. **EUROPE**

10. **SCANDINAVIE**
   Denmark
   Finlande
   Norvège
   Suède

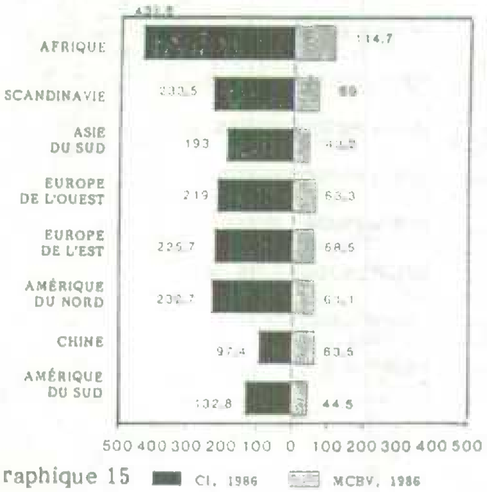
11. **EUROPE DE L'EST**
   Tchécoslovaquie
   Hongrie
   Pologne
   Roumanie
   URSS
   Yougoslavie

Taux de mortalité chez les Canadiens de 35 ans et plus, normalisés selon l'âge et groupés sur 5 ans, pour les cardiopathies ischémiques et et les maladies cérébrovasculaires, 1969-1973



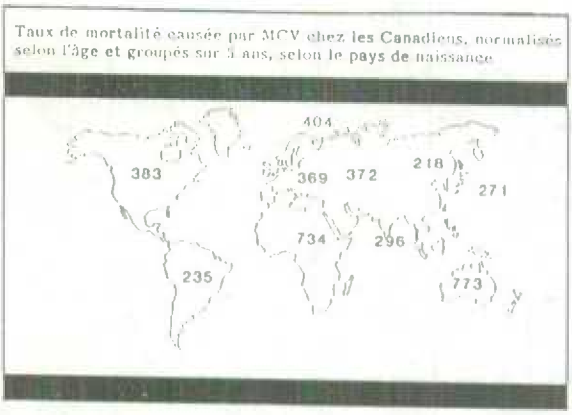
Graphique 14

Taux de mortalité chez les Canadiens de 35 ans et plus, normalisés selon l'âge et groupés sur 5 ans, pour les cardiopathies ischémiques et les maladies cardio-vasculaires, 1984-1988



Graphique 15

Le graphique 16 transpose les taux de mortalité causée par MCV des Canadiens sur une carte du monde. Dans l'ensemble, les taux sont faibles pour les Canadiens d'origine sud-américaine, chinoise ou sud-asiatique de la première génération; ils sont élevés dans le cas de la Scandinavie et de l'Afrique; les taux observés pour l'Amérique du Nord se situent au milieu et ressemblent à ceux de l'Europe de l'est et de l'eurape de l'Ouest. Cet ordre général des taux de mortalité causée par MCV chez immigrants canadiens est comparé à celui constaté par l'OMS (Organisation Mondiale de la Santé, 1988) selon le pays de naissance (tableaux 4 et 5). Autant que l'on puisse en juger, les taux suivent une même évolution dans les deux cas. Le manque de données détaillées empêche d'établir une comparaison rigoureuse entre les TMNA chez les Canadiens de la première génération et ceux observés dans leur pays de naissance.


Taux de mortalité causée par MCV chez les Canadiens, normalisés selon l'âge et groupés sur 5 ans, selon le pays de naissance
404
383
369
372
218
271
734
296
235
773

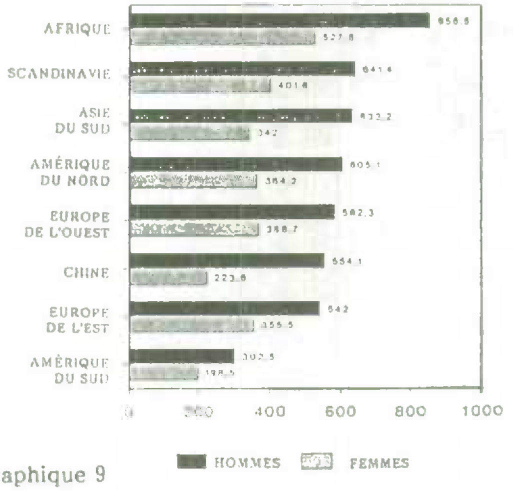
Graphique 16

4. DISCUSSION

La mortalité causée par MCV est en régression depuis 35 ans, mais elle continue d'occuper le premier rang au Canada. Les MCV comptent pour plus de 40% de l'ensemble des décès. Le taux de mortalité causée par MCV ont tendance à être plus élevés chez les hommes que chez les femmes, et il en est de même pour les taux de mortalité des Canadiens de la première génération, normalisés selon l'âge et groupés sur 5 ans. Dans ce dernier cas, toutefois, les taux varient beaucoup selon le pays de naissance. Ils sont élevés pour la Scandinavie et l'Afrique; faibles pour l'Amérique de Sud, la Chine et l'Asie du Sud; et en baisse pour l'ensemble de la population et les Canadiens de la première génération, sauf ceux de l'Afrique (35 ans et plus). L'OMS rapporte que les taux de mortalité causée par MCV diminuent à l'échelle mondiale excepté pour les pays de l'Europe de l'Est (Organisation Mondiale de la Santé, 1988). Nous n'avons pas trouvé de hausse de ces taux pour les Canadiens originaires de l'Europe de l'Est. Ceci peut s'expliquer par les critères de sélection lors du processus d'immigration ainsi que par un régime alimentaire différent. Dans les cas des pays de l'Afrique, il n'existe pas de données comparables mais, selon une récente communication de Balfe *et al.*, 1988, les taux de mortalité causée par MCV sont très élevés chez les Blancs de l'Afrique du Sud et chez les Asiatiques. Ces deux groupes constituent la majorité des Canadiens de la première génération nés en Afrique.

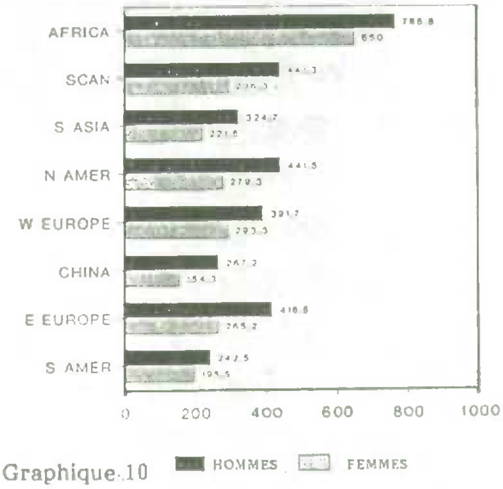
Dans un article publié récemment sur la mortalité causée par MCVchez les Sud-asiatiques habitant l'Asie ou ailleurs, McKeigue *et al.*, 1989, signalent que les taux de mortalité causée par MCV chez les Sud-asiatiques sont parmi les plus élevés du monde. L'article traite aussi de leur prédisposition au diabète sucré, notamment celui qui n'est pas insulino-dépendant. Le rapport entre ces deux affections semble manifeste, (Salonem, 1989;

Taux de mortalité causée par MCV chez les 35 ans et plus au Canada, normalisés selon l'âge et groupés sur 5 ans, d'apres le pays de naissance et le sexe, 1969-1973



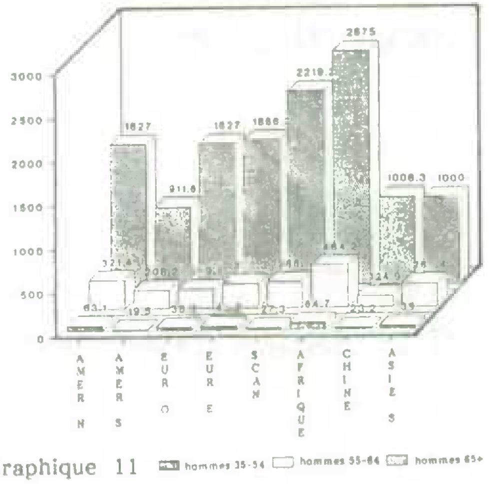
Graphique 9

Taux de mortalité causée par MCV chez les 35 ans et plus au Canada, normalisés selon l'âge et groupés sur 5 ans, d'après le pays de naissance et le sexe, 1984-1988



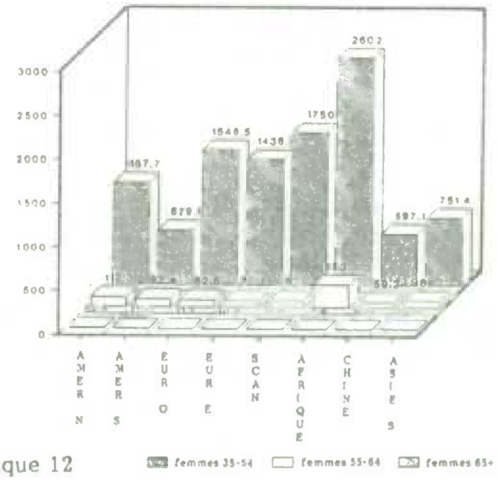
Graphique 10

Taux de mortalité causée par MCV chez les hommes de 35 ans et plus, groupés sur 5 ans, par âge et d'après le pays de naissance, 1984-1988



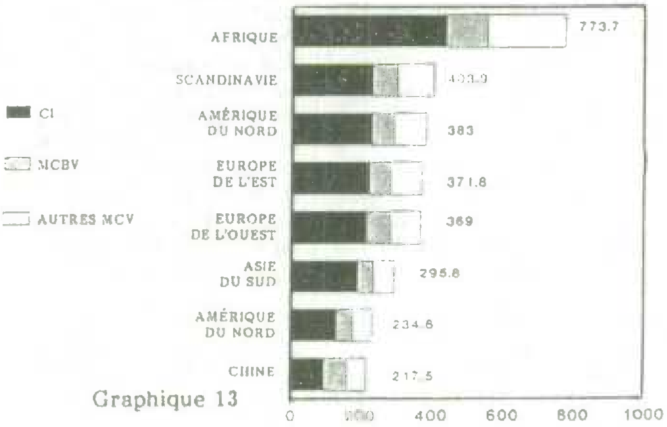
Graphique 11

Taux de mortalité causée par MCV chez les femmes de 35 ans et plus, groupés sur 5 ans, par âge et d'après le pays de naissance, 1984-1988



Graphique 12

Taux de mortalité chez Canadiens de 35 ans et plus, normalisés selon l'âge et groupés sur 5 ans, pour les cardiopathies schémiques, les maladies cérébrovasculaires et d'autres maladies cardio-vasculaires, 1984-1988

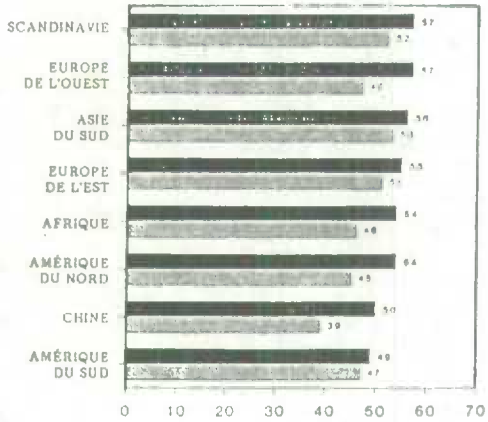


Graphique 13

Le graphique 7 montre les TMNA au Canada, d'après le pays de naissance, pour l'ensemble des causes et les deux intervalles. Pour l'ensemble des causes, les taux de mortalité entre 1984-88 varient selon le pays de naissance: les derniers en rang sont L'Amérique du Sud, la Chine et l'Asie du Sud, tandis que l'Afrique vient en tête. Les TMNA sont en régression entre 1969-1973 et 1984-1988 quelle que soit la provenance, sauf pour l'Afrique.

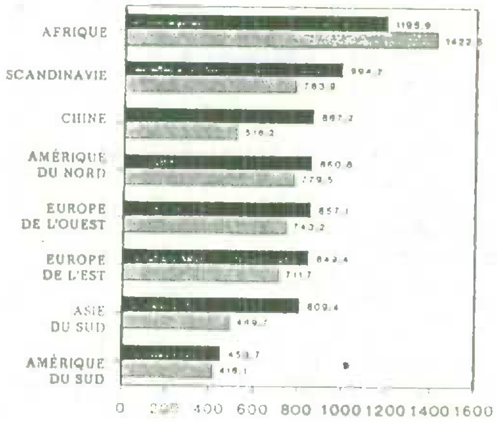
Le graphique 8 montre les taux de mortalité causée par MCV normalisés selon l'âge au Canada, chez les 35 ans et plus, d'après le pays de naissance. Eux aussi varient selon le pays de naissance et leur évolution ressemble à celle des taux observés pour l'ensemble des causes. Ils sont en baisse, sauf pour l'Afrique (35 ans et plus), et la régression est la plus forte pour la Chine et l'Asie du Sud. Les pays de naissance suivent un ordre similaire pour les deux intervalles.

Taux de mortalité causée par MCV au Canada chez les 35 ans et plus, (1969-73 et 1984-88)



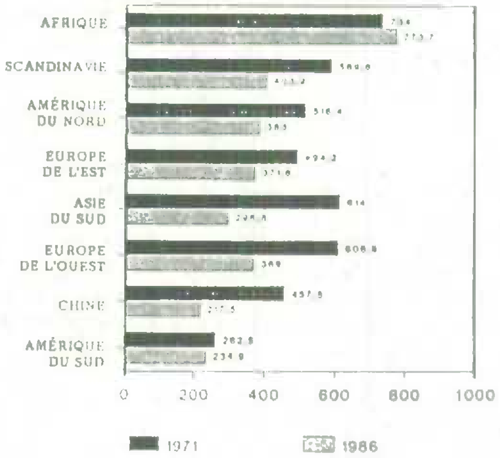
Graphique 6

Taux de mortalité chez 35 ans et plus pour l'ensemble des causes au Canada, normalisés selon l'âge et groupés sur 5 ans, d'après le pays de naissance, 1971 et 1986



Graphique 7

Taux de mortalité causée par MCV chez les 35 ans et plus au Canada, normalisés selon l'âge et groupés sur 5 ans, d'après le pays de naissanee



Graphique 8

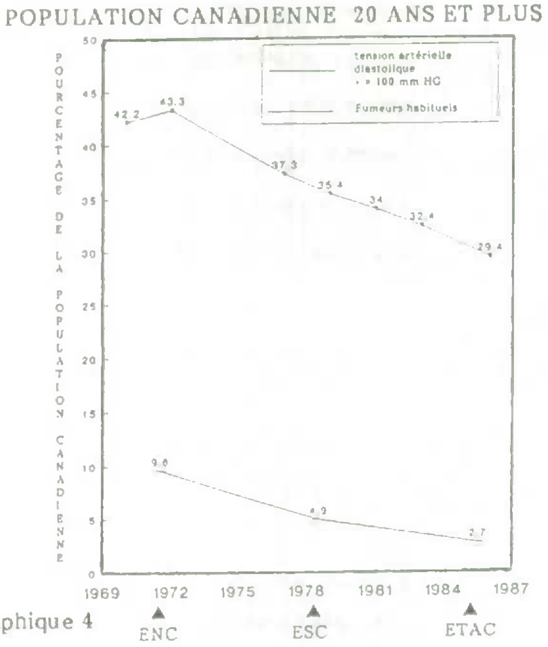
Les graphiques 9 et 10 montrent les différents taux de mortalité causée par MCV chez les hommes et chez les femmes pour les deux intervalles. Ils sont régulièrement plus élevés chez les hommes que chez les femmes, quel que soit le pays de naissance. Ils sont beaucoup moins élevés chez les femmes d'origine chinoise que chez les hommes d'origine chinoise. Le rang observé selon le pays de naissance est à peu près le même chez les hommes et chez les femmes.

Les graphiques 11 et 12 indiquent, pour les hommes et les femmes respectivement, les taux de mortalité causée par MCV par âge et d'après le pays de naissance, pour l'intervalle 1984-1988. Les taux montent en flèche selon l'âge. L'évolution est à peu près la même selon l'âge et le sexe.

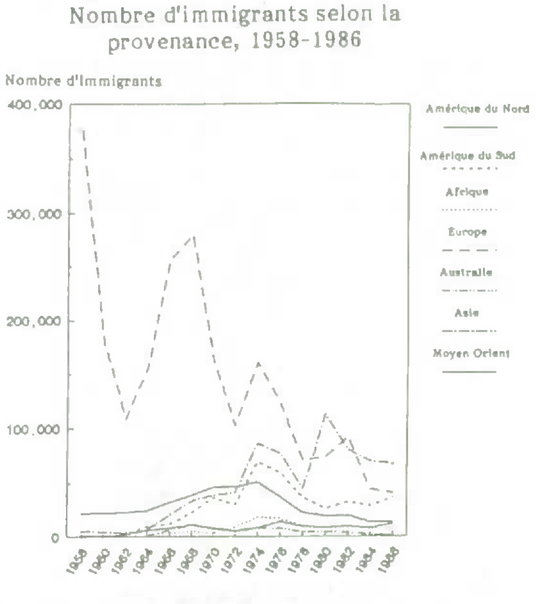
Le graphique 13 montre, pour 1984-1988 et d'après le pays de naissance, les TMNA au Canada en ce qui concerne les cardiopathies ischémiques (CI), les maladies cérébrovasculaires (MCBV) et d'autres MCV. Les CI comptent pour plus de 50 % de l'ensemble des décès attribuables aux MCV. La part des MCBV est bien moins élevée, mais elle demeure tout de même importante. À noter que les MCBV représentent, chez les personnes d'origine chinoise, le tiers de tous les décès causés par MCV.

Les graphiques 14 et 15 indiquent, pour 1969-1973 et 1984-1988 respectivement, les TMNA au Canada pour les CI et les MCBV. Les décès par CI sont en baisse pour la Chine et l'Asie du Sud et ils suivent à peu près la même évolution d'après le pays de naissance.

tendance contribue à enrichir la mosaïque multiculturelle et multi-ethnique qui caractérise la population canadienne. Il a été démontré par ailleurs que les taux de mortalité par MCV varient selon la provenance (Thom, 1988). Les études épidémiologiques des Canadiens de la première génération, selon le lieu de naissance, peuvent donc servir à évaluer l'état de santé de la population et à planifier des programmes d'intervention en matière de santé. Le présent document examine les taux de mortalité causée par MCV chez les immigrants canadiens, selon la provenance.



Graphique 4



Graphique 5

(1) Y compris le Mexique, l'Amérique centrale et les Caraïbes.
(2) Y compris l'Océanie.

2. MÉTHODOLOGIE

La présente étude s'appuie sur l'information tirée du recensement de la population du Canada (Statistique Canada, 1971 et 1986) et de la base canadienne de données sur la mortalité (BCDM) (Statistique Canada). Aux fins de l'étude et pour les besoins de représentation des différences ethniques, les pays de naissance ont été groupés géographiquement comme l'indique le tableau 1. Le nombre de pays choisis pour l'étude a été restreint par les données de la BCDM. Par exemple, on a fini par exclure le Japon et l'Australie de la plus grande partie de l'analyse en raison des faibles taux. Les périodes considérées ont aussi été limitées par les données disponibles (tableau 2). On a choisi deux intervalles de cinq ans, 1969-1973 et 1984-1988, parce que les données étaient les plus complètes.

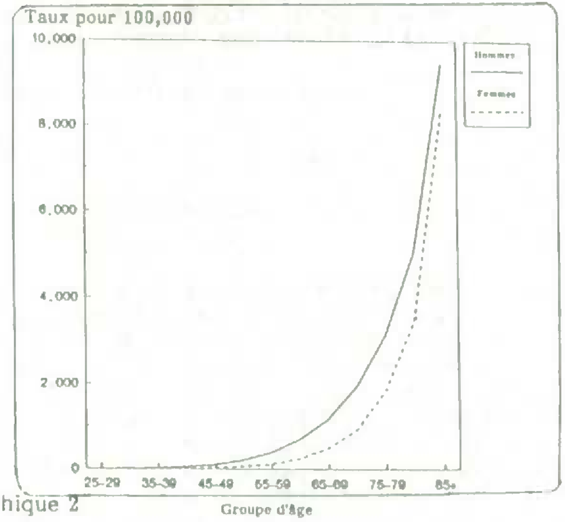
Le taux de mortalité causée par MCV en Alberta est connu pour 1984, mais l'absence de données sur le pays de naissance oblige à faire une estimation. Le nombre de décès au Canada en 1988 est estimé d'après celui observé en 1987. Toutes les estimations ont fait l'objet d'un examen destiné à vérifier si elles concordent avec les tendances relevées.

Le choix d'intervalles de cinq ans vise à obtenir un nombre de décès raisonnable dans les groupes de pays retenus. Les données des recensementd de 1971 et de 1986 -- correspondant aux points milieux des intervalles de cinq ans -- sont classées selon le pays de naissance, l'âge (35-54, 55-64, 65+) et le sexe. Les données tirées de la BCDM sont aussi groupées selon l'âge, le sexe et le pays de naissance et, de plus, selon la cause de décès (ensemble des causes, MCV, cardiopathies ischémiques et maladies cérébrovasculaires). Les données du recensement servent de dénominateur et la somme des décès sur chaque intervalle sert de numérateur, ce qui fournit les taux de mortalité groupés sur cinq ans. Les taux de mortalité selon l'âge sont normalisés pour la population canadienne de 1986. Lorsqu'il en est fait mention ci-après, il s'agit des données groupées sur cinq ans et normalisées selon l'âge pour les trois groupes indiqués plus haut, à partir de 35 ans.

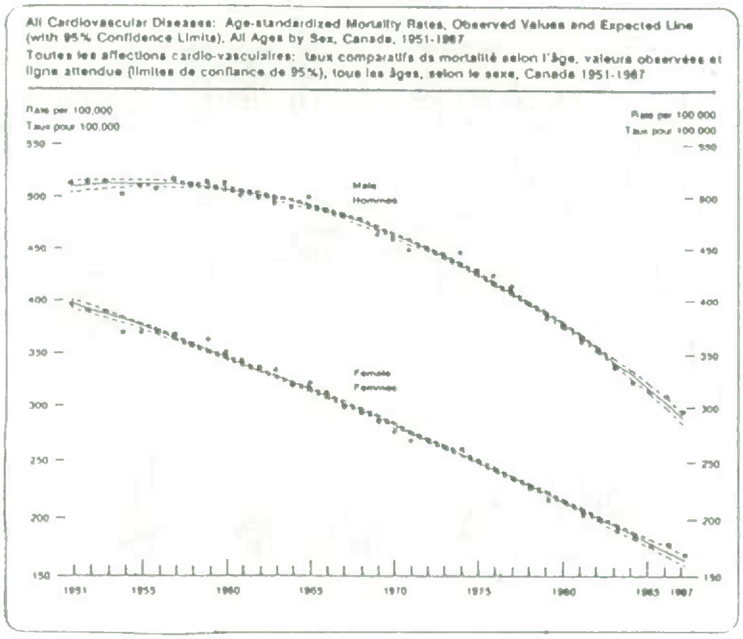
3. RÉSULTATS

Le graphique 6 montre le ratio de la mortalité causée par MCV à la mortalité pour l'ensemble des causes, selon le pays de naissance, pour les intervalles 1969-1973 et 1984-1988, respectivement. Entre 1969 et 1973, les MCV entraînent au moins 50 % de tous les décès chez les plus de 35 ans et elles dominent dans les taux de mortalité des Sud-asiatiques. Entre 1984 et 1988, les taux de mortalité causée par MCV tombent à 40-50 % pour la plupart des pays de naissance.

Maladies Cardio-Vasculaires
Taux de mortalité pour 100,000,
Canada, 1987



Graphique 2


All Cardiovascular Diseases: Age-standardized Mortality Rates, Observed Values and Expected Line (with 95% Confidence Limits), All Ages by Sex, Canada, 1951-1987

Toutes les affections cardio-vasculaires: taux comparatifs de mortalité selon l'âge, valeurs observées et ligne attendue (limites de confiance de 95%), tous les âges, selon le sexe, Canada 1951-1987

Graphique 3

Bonne nouvelle, cependant: les TMNA sont en régression depuis 35 ans. S'ils s'étaient maintenus à leur maximum, ils auraient causé encore 22,000 décès chez les hommes et encore 13,000 chez les femmes.

Trois principaux facteurs de risque de MCV ont été cernés: le tabagisme, l'hypertension artérielle et l'hypercholestérolémie. Pour les deux premiers, le présent document établit une comparaison des données selon deux intervalles. Le graphique 4 indique que la proportion de fumeurs au Canada est tombée de 42.2 % en 1970 à 29.4 % en 1986 (Santé et Bien-être, 1973). Il montre aussi la baisse dans la proportion de Canadiens dont la tension artérielle diastolique est égale ou supérieure à 100 mm HG, c'est-à-dire de 9.6 % selon l'enquête Nutrition Canada (Santé et Bien-être, 1973), à 4.9 % selon l'enquête Santé Canada (Santé et Bien-être, 1981) et à 2.7 % selon l'enquête sur la tension artérielle des Canadiens (Santé et Bien-être, 1989). Cette baisse peut expliquer en partie la régression des TMNA depuis les 30 dernières années.

Environ un résident canadien sur six est actuellement un immigrant de la première génération (Emploi et Immigration, 1988). Ces dernières années, la provenance des immigrants tend à évoluer des pays de l'Europe à ceux d'autres parties du monde. Le graphique 5 illustre l'évolution temporelle, selon l'origine des immigrants. Les taux d'immigration sont en baisse pour l'Europe, mais en hausse pour l'Asie et l'Amérique du Sud. Cette

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

L'ÉVOLUTION DES CARACTÉRISTIQUES DÉMOGRAPHIQUES ET LES MALADIES CARDIO-VASCULAIRES

H. Johansen[1], C. Nair[2], M. Nargundkar[3] et J. Strachan[4]

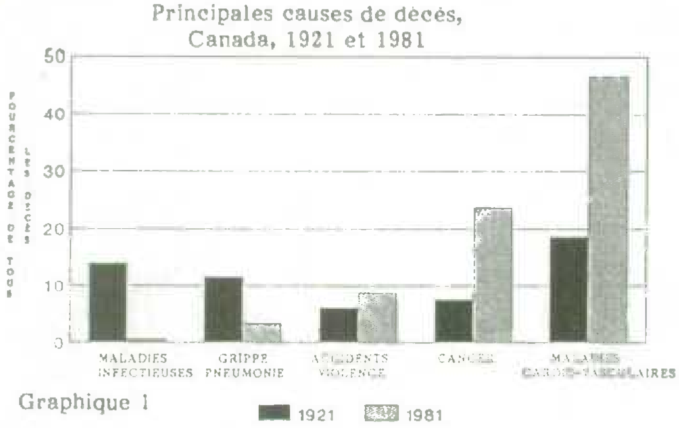
RÉSUMÉ

MOTS CLÉS: maladies cardio-vasculaires, taux de mortalité, immigrants, origine ethnique, lieu de naissance, tendances temporelles.

Les maladies cardio-vasculaires (MCV) sont la principale cause de mortalité au Canada tout comme dans la plupart des pays industrialisés. Les facteurs de risque sont principalement le tabagisme, l'hypertension artérielle et l'hypercholestérolémie. Environ un résident canadien sur six est un immigrant de la première génération. Il a été démontré que les taux de mortalité causée par MCV varient selon l'origine ethnique et diffèrent chez les groupes de la première génération au Canada. Dans l'ensemble, ils sont faibles chez les groupes d'origine sud-américaine, chinoise ou sud-asiatique, alors qu'ils sont élevés chez ceux d'origine scandinave ou africaine. Les taux observés pour l'Amérique du Nord sont similaires à ceux constatés pour l'Europe de l'Est et l'Europe de l'Ouest. Entre deux intervalles de cinq ans, soit 1969-1973 et 1984-1988, ils s'inscrivent généralement en baisse, sauf chez les groupes venus de l'Afrique (35 ans et plus). Ils suivent un ordre similaire selon l'âge et le sexe et ils sont régulièrement plus élevés chez les hommes que chez les femmes.

1. INTRODUCTION

Les maladies cardio-vasculaires (MCV) constituent la principale cause de décès au Canada tout comme dans la plupart des pays industrialisés. Elles sont aussi la principale cause d'hospitalisation au Canada: on estime à plus de $3 milliards par année les frais hospitaliers qu'elles entraînent directement (Nair *et al.*, 1989). En 1987, elles ont causé la mort de plus de 77,000 Canadiens (Nair *et al.*, 1989), ou l'équivalent de la population de Kingston, en Ontario. Ce n'est pas depuis hier qu'elles sont la principale cause de décès au Canada. En 1921, elles ont occasionné 18.6 % de l'ensemble des décès; en 1981, elles venaient toujours en tête, comptant pour 46.6 % de tous les décès (graphique 1).



Graphique 1

En 1987, les MCV ont causé 43 % de l'ensemble des décès, ce qui implique que 4 Canadiens sur 10 risquent d'en mourir chaque année. Le graphique 2 montre les taux de mortalité causée par MCV en 1987, selon l'âge et le sexe. Il indique que les décès surviennent plus tôt chez les hommes et que, dans tous les groupes d'âge, les taux sont plus élevés chez les hommes que chez les femmes. Comme l'on doit s'y attendre, ils sont beaucoup plus élevés chez les personnes de plus de 65 ans que chez celles qui sont moins âgées. Ils sont environ quatre fois supérieurs chez celles de 75 ans et plus que chez celles de moins de 65 ans. D'où les répercussions majeures qu'entraine sur les services de santé le vieillissement de la population. Le graphique 3 montre les taux de mortalité normalisés selon l'âge (TMNA) et le sexe, de 1951 à 1987.Bonne nouvelle, cependant: les TMNA sont en régression depuis 35 ans. S'ils s'étaient maintenus à leur maximum, ils auraient causé encore 22,000 décès chez les hommes et encore 13,000 chez les femmes.

[1] H. Johansen, Direction de la promotion de la santé, Santé et Bien-être social Canada, Ottawa, (Ontario) K1A 1B4.

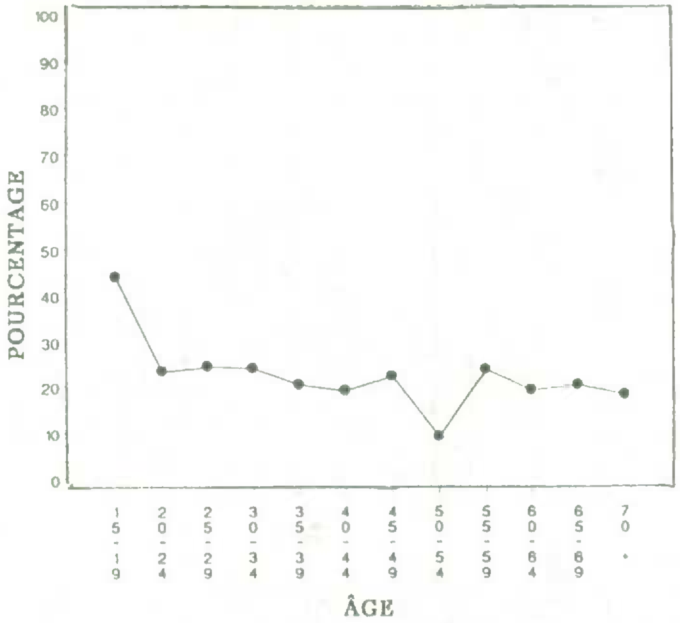
[2] C. Nair, Centre canadien d'information sur la santé, Statistique Canada, Ottawa, (Ontario) K1A 1B4.

[3] M. Nargundkar, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, (Ontario) K1A 1B4.

[4] J. Strachan, School of Health Information Science, University of Victoria, Victoria, Columbie-Britannique, V8W 2Y2.

FIGURE 2. POURCENTAGE D'INDIVIDUS N'AYANT JAMAIS FUMÉ, SELON 12 GROUPES D'ÂGE

A. HOMMES



B. FEMMES

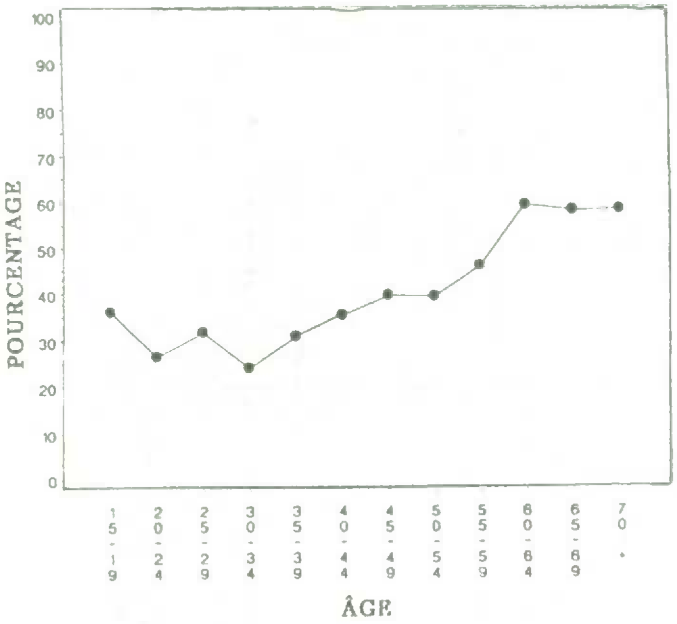
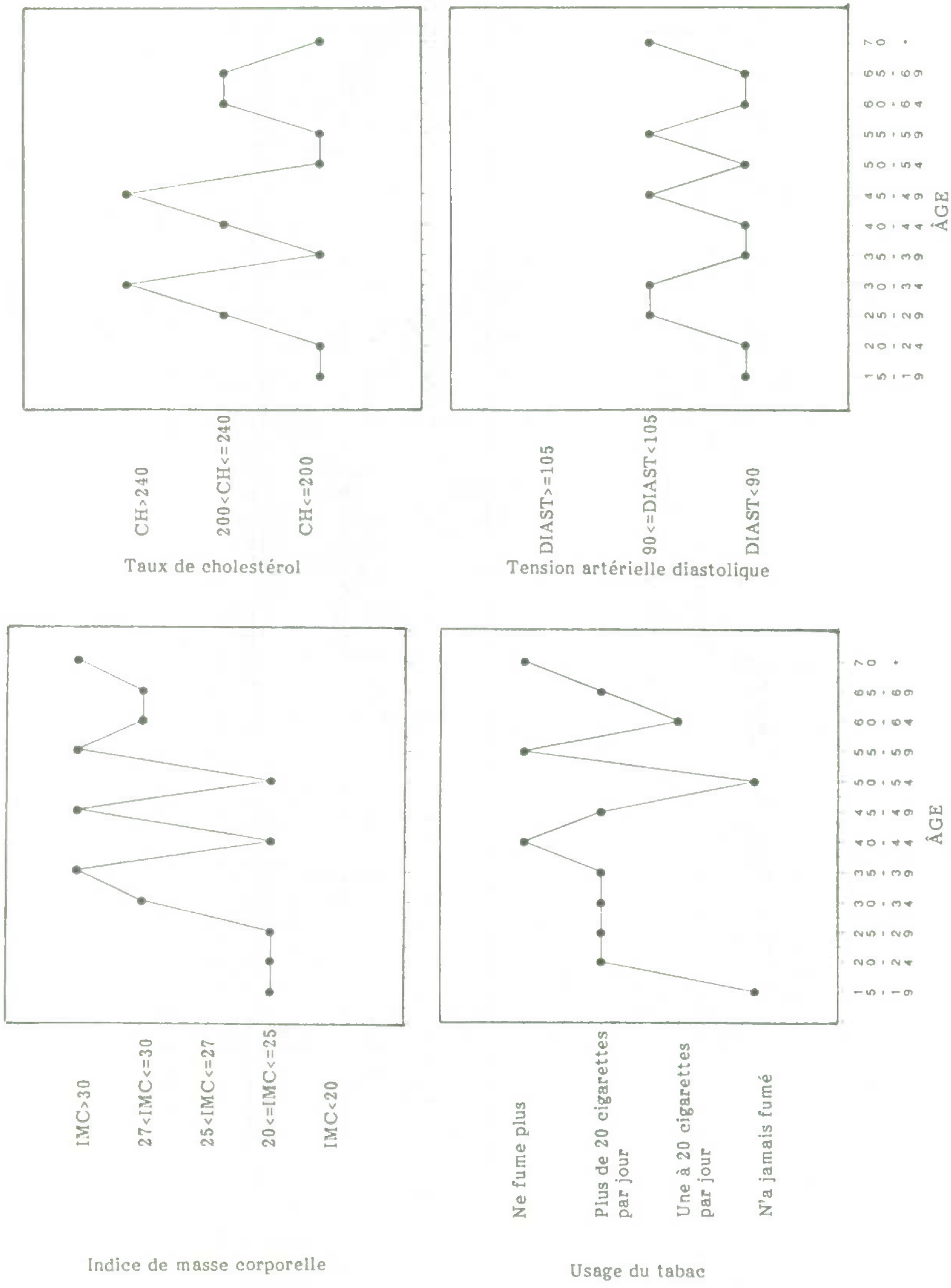


FIGURE 1. (suite)

B. IRENÉE (cycle irrégulier)



Taux de cholestérol

Tension artérielle diastolique

Indice de masse corporelle

Usage du tabac

FIGURE 1. CYCLES DE VIE SIMULÉS

A. RÉGIS (cycle régulier)

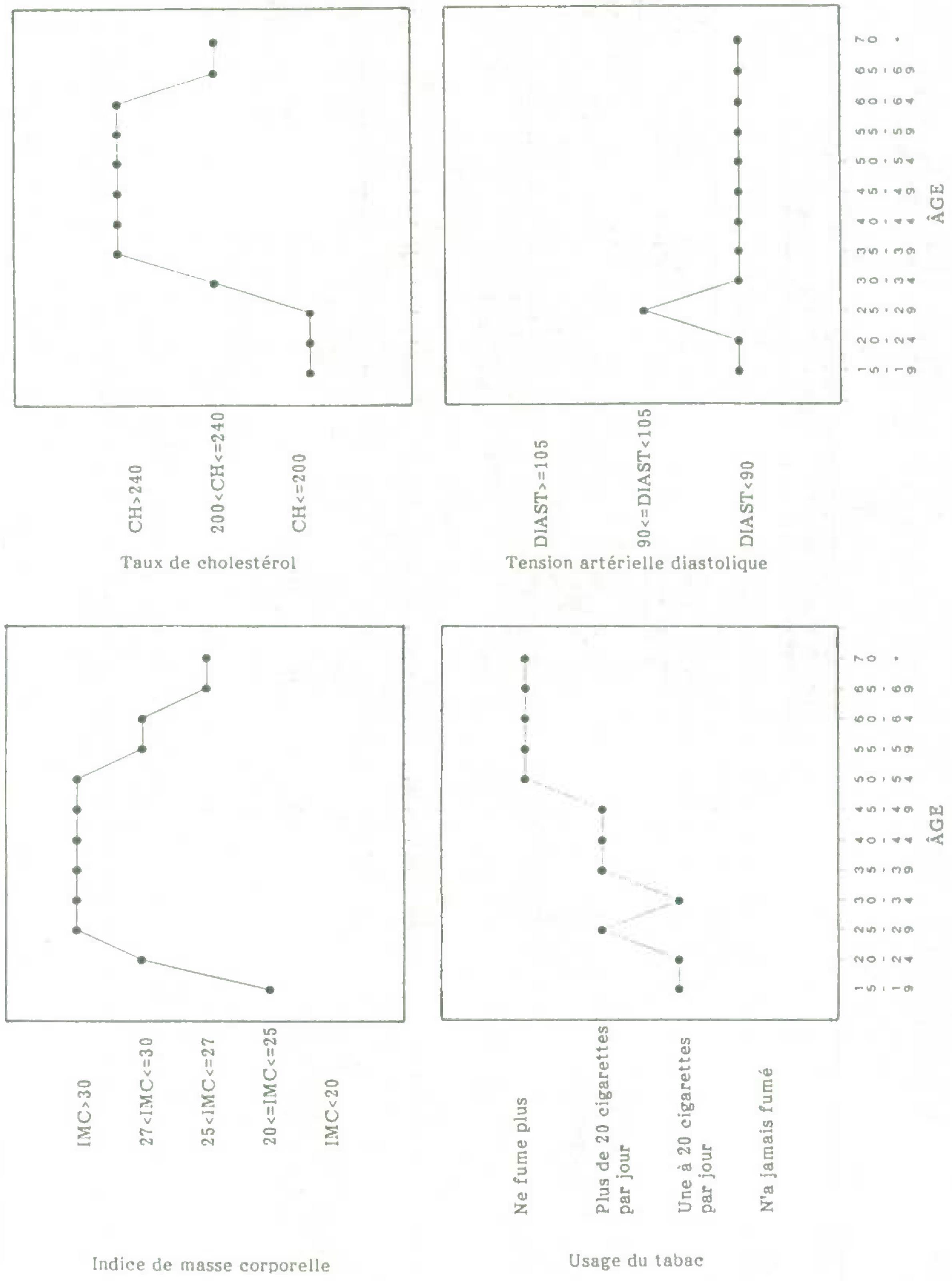


TABLEAU 1

B. Probabilités de transition pour la variable 4
(usage du tabac); hommes des groupes d'âge 30-34 et 35-39
les éléments (1,4), (2,1), (3,1) et (4,1) sont obligatoirement nuls

		Wij = |i-j|			
		N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus
Z	N'a jamais fumé	.98	.02	.00	.00
	1 à 20 cigarettes par jour	.00	.75	.17	.08
	Plus de 20 cigarettes par jour	.00	.00	1.00	.00
	Ne fume plus	.00	.00	.00	1.00

	Wij = (i-j)²			
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus
Z	.98	.02	.00	.00
	.00	.75	.25	.00
	.00	.00	.93	.07
	.00	.00	.00	1.00

		Wij = |i-j|			
		N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus
Z'	N'a jamais fumé	.98	.02	.02	.00
	1 à 20 cigarettes par jour	.00	.78	.22	.00
	Plus de 20 cigarettes par jour	.00	.00	93	.07
	Ne fume plus	.00	.00	.00	1.00

	Wij = (i-j)²			
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus
Z'	.98	.02	.00	.00
	.00	.75	.25	.00
	.00	.00	.93	.07
	.00	.00	.00	1.00

TABLEAU 1

A. Fréquences de transition pour la variable 4
(usage du tabac), hommes des groupes d'âge 30-34 et 35-39
les éléments (1,4), (2,1), (3,1) et (4,1) sont obligatoirement nuls

	Wij = |i-j|				
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus	
Z N'a jamais fume	152,389	3,263	0	0	155,652
1 à 20 cigarettes par jour	0	111,080	24,390	11,661	147,131
Plus de 20 cigarettes par jour	0	0	161,514	0	161,514
Ne fume plus	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

	Wij = (i-j)²				
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus	
Z	152,389	3,263	0	0	155,652
	0	111,080	36,051	0	147,131
	0	0	149,853	11,661	161,514
	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

	Wij = |i-j|				
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus	
Z' N'a jamais fumé	152,389	0	3,263	0	155,652
1 à 20 cigarettes par jour	0	114,343	32,788	0	147,131
Plus de 20 cigarettes par jour	0	0	149,853	11,661	161,514
Ne fume plus	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

	Wij = (i-j)²				
	N'a jamais fumé	1 à 20 cigarettes par jour	Plus de 20 cigarettes par jour	Ne fume plus	
Z'	152,389	3,263	0	0	155,652
	0	111,080	36,051	0	147,131
	0	0	149,853	11,661	161,514
	0	0	0	222,417	222,417
	152,389	114,343	185,904	234,078	686,714

groupes d'âge. La technique de lissage a donc aussi pour avantage de révéler certains types d'incohérences dans les données de départ.

OBSERVATIONS FINALES

Hétérogénéité, sélection et tables de mortalité multidimensionnelles sont trois concepts statistiques qui se rejoignent lorsqu'on considère les problèmes causés par l'utilisation de données transversales au lieu de données longitudinales. Des recherches additionnelles seront nécessaires pour savoir comment discerner et résoudre ces problèmes et il faudra aussi plus de données longitudinales pour les éviter.

Sur le plan informatique, en dépit de la quantité parfois prohibitive de ressources qu'exigent les modèles de microsimulation, il y a tout lieu de croire que de nouveaux progrès technologiques, comme l'augmentation de la vitesse de traitement et de la capacité de mémoire, l'utilisation plus intensive d'ordinateurs spécialisés et le traitement en parallèle, viendront accroître les possibilités des modèles de microsimulation (voir Hoschka (1986)).

Les auteurs tiennent à remercier Michael Wolfson pour les avoir incités à réaliser cette étude, dont les résultats serviront au modèle de microsimulation de la santé POHEM (voir Wolfson (1989)). Ils tiennent aussi à exprimer leur reconnaissance à Monica Tomiak pour sa précieuse collaboration sur les plans technique et informatique.

BIBLIOGRAPHIE

Barr, R.S. and Turner, J.S. (1981). Microdata File Merging Through Large Scale Network Technology. Math. Prog. Study, **15**, 1-22.

Hitchcock, F.L. (1941). The Distribution of a Product from Several Sources to Numerous Localities. J. Mathematical Physics, **10**, 224-230.

Hoschka, Peter (1986). Requisite Research on Methods and Tools for Microanalytic Simulation Models. In Orcutt, et al. (1986), pp. 45-54.

Lawler, Eugene L. (1976). *Combinatorial Optimization: Networks and Matroids.* Holt Rinehart and Winston.

Orcutt, Guy; Merz, Joachim; and Quinke, Hermann (1986), Editors. *Microanalytic Simulation Models to Support Social and Financial Policy.* Proceedings of 1983 symposium in Bonn, Germany. North-Holland.

Rogers, Andrei (editor) (1980). Essays in Multistate Mathematical Demography. Special issue of Environment and Planning A 12(5).

SAS Institute Inc. (1985). *SAS/OR User's Guide*, Version 5 Edition.

SAS Institute Inc. (1986). Technical Report: P-146. Changes and Enhancements to the Version 5 SAS System.

Statistics Canada and National Health and Welfare (1981). The Health of Canadians. Report of the Canada Health Survey. Catalogue 82-538E. Ottawa, Canada.

Wolfson, Michael C. (1989). A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data. Paper presented at 21st General Conference of the International Associaiton for Research in Income and Wealth, Lahnstein, West Germany, Aug. 20-26, 1989.

Wolfson, Michael and Birkett, Nick (1989). POHEM, Population Health Module of the System of Health Statistics (SHS): Preliminary Exporation of CHD. Paper presented at Canadian Epidemiology Research Conference, Ottawa, August 1989.

(1941)). Les contraintes fondamentales imposent des totaux connus pour les sommes sur les rangées et colonnes des éléments non négatifs d'une matrice. En classant les états multidimensionnels dans un ordre approprié, nous pouvons imaginer une matrice qui aurait pour éléments des fréquences de transition et où les rangées correspondraient à l'état de départ et les colonnes, à l'état d'arrivée. Le nombre d'individus qui passent d'un état à un di autre est manifestement non négatif et les totaux de rangée et de colonne correspondent au nombre total d'invidus qui se trouvent dans les états de départ et les états d'arrivée respectivement. Nous voulons faire en sorte que ces totaux concordent avec les éléments de la matrice, ce qui en fait un problème de transport. Dans le contexte de notre étude, la fonction économique attribue des coûts à diverses transitions, un peu comme on le fait dans l'application classique, lorsqu'on attribue des coûts d'expédition aux divers trajets possibles. Les coefficients sont tels que les transitions qui s'opèrent entre états voisins sont moins coûteuses que celles qui s'opèrent entre des états éloignés l'un de l'autre.

Le fait de reconnaître que la question qui nous occupe relève de la planification par réseaux a des conséquences importantes au point de vue théorique et pratique. Une propriété intéressante est que les solutions sont en nombres entiers. Si les totaux de rangée et de colonne sont des entiers, les algorithmes donneront des solutions optimales en nombres entiers de sorte que le nombre d'individus qui passent d'un état à un autre ne sera jamais fractionnaire (voir Lawler (1976)). Un programme linéaire général produit difficilement des solutions optimales en nombres entiers mais la planification par réseaux est conçue pour cela.

Autre avantage intéressant: la résolution des problèmes de planification par réseaux se fait beaucoup plus rapidement et exige moins d'espace en mémoire que les programmes linéaires généraux. Avec moins de 400 noeuds, les problèmes que nous avons pu résoudre dans cet article ne peuvent être qualifiés de majeurs selon les normes du domaine et ils sont couramment résolus à l'aide des programmes disponibles. Dans une application d'envergure (voir Barr et Turner (1981)), on a pu résoudre un problème de transport qui comportait plus de 20,000 contraintes et 10,000,000 de variables.

Dans la version la plus élémentaire d'un problème de transport, toutes les transitions sont permises et il n'existe aucune limite supérieure pour les fréquences. Dans ces conditions, et dans la mesure où les totaux de rangée et de colonne concordent entre eux (c.-à-d. que le total général est le même dans les deux sens), le problème a toujours une solution (c.-à-d. qu'il est "réalisable" en termes de programmation mathématique). Le modèle analysé a plus de souplesse que veut le laisser à entendre cette formulation élémentaire et cette souplesse est nécessaire dans le cadre de notre analyse. Par exemple, si certaines transitions sont inconcevables, on peut les exclure du modèle. On peut aussi définir des limites inférieures et supérieures pour diverses variables dans la solution. Cela revient à restreindre l'intervalle des valeurs de certaines probabilités de transition en fonction de ce que l'on croit être vraisemblable et invraisemblable. Lorsqu'on ajoute des contraintes de ce genre, on risque de rendre le problème irréalisable. (Par exemple, si un nombre suffisant de transitions sont exclues du modèle, le programme risque de ne pas pouvoir satisfaire la demande à certains noeuds.) Avec les données dont nous disposions, il est arrivé que nous ne trouvions pas de solution à un problème et la cause de cette impasse était toujours la même: dans le cas de l'usage du tabac, nous avions établi qu'il était impossible pour un individu d'appartenir à la classe "n'a jamais fumé" après avoir appartenu à l'une ou l'autre des trois autres classes (ce qui était une contrainte acceptable) mais après avoir rajusté les données brutes pour réaliser la concordance des totaux, nous avons constaté qu'il y avait plus d'individus dans la catégorie "n'a jamais fumé" à l'âge t+1 qu'à l'âge t. Cette incohérence n'était pas due à un vice d'application de la méthode mais au fait qu'on avait utilisé des données transversales à la place de données longitudinales. En un sens, les données rajustées renferment des valeurs aberrantes (valeurs incompatibles avec le modèle) et doivent être redressées pour faire en sorte que la proportion d'individus n'ayant jamais fumé n'augmente jamais d'un groupe d'âge à l'autre.

Nous avons utilisé deux procédures SAS pour notre analyse: LP pour les programmes linéaires généraux (SAS Institute (1985)) et NETFLOW pour la planification par réseaux (SAS Institute (1985)). Une troisième procédure, TRANS (SAS Institute (1985)), est conçue spécialement pour les problèmes de transport; cependant, des erreurs dans le programme, corrigées depuis par SAS, nous ont obligé à renoncer à cette procédure. Un logiciel supérieur pour la planification par réseaux, TNETFLOW (SAS Institute (1986)), est maintenant disponible.

La figure 2A donne le pourcentage d'individus n'ayant jamais fumé (hommes seulement) pour chacun des 12 groupes d'âge (selon l'Enquête Santé Canada). Si l'on fait exception du pourcentage élevé observé pour la première tranche d'âge et du pourcentage relativement faible observé pour la tranche d'âge 50-54, les observations décrivent une courbe approximativement linéaire qui a une pente négative mais qui ne décroît pas uniformément. Afin d'ajuster ces données, nous avons exécuté une régression linéaire simple du logarithme de la proportion d'individus n'ayant jamais fumé par rapport à l'âge en ne tenant pas compte des deux groupes d'âges précités. Parfois, il a fallu remplacer des observations par les valeurs ajustées correspondantes pour obtenir une proportion non croissante d'individus n'ayant jamais fumé. Les proportions d'individus pour les autres classes de la variable ont conservé leur importance relative.

La figure 2B donne le pourcentage de femmes qui n'ont jamais fumé pour chacun des 12 groupes d'âge (selon l'Enquête Santé Canada). Les observations révèlent une nette tendance à la hausse qui serait fort probablement attribuable à un effet de cohorte contenu dans les données transversales: en 1978, les femmes relativement plus âgées étaient plus susceptibles d'appartenir à la catégorie des personnes n'ayant jamais fumé que les femmes relativement moins âgées. Cela montre bien les difficultés auxquelles on s'expose lorsqu'on utilise des données transversales au lieu de données longitudinales. Si nous appliquions nonchalamment les méthodes ci-dessus aux données transversales, nous n'obtiendrions pas de solution pour la plupart des transitions entre individu: n'a jamais fumé -- fume de 1 à 20 cigarettes -- fume plus de 20 cigarettes - ne fume plus. Nous avons appliqué la méthode PL en nous servant de diverses combinaisons de poids (valeur absolue de la distance oindividu: n'a jamais fumé -- fume de 1 à 20 cigarettes -- fume plus de 20 cigarettes - ne fume plus. Nous avons appliqué la méthode PL en nous servant de diverses combinaisons de poids (valeur absolue de la distance ou carré de la distance) et de fonctions économiques (z ou z'). Dans les quatre cas, les éléments (1,4), (2,1), (3,1) et (4,1) des matrices de transition (qui représentent respectivement les cas où un individu passerait de l'état (i) à l'état (iv) et de l'état (ii), (iii) ou (iv) à l'état (i)) sont obligatoirement nuls tandis qu'aucune restriction n'est imposée pour les individus qui cessent de fumer ou recommencent à fumer.

Selon le tableau 1, le changement de fonction économique (de z à z') entraîne une modification des fréquences de transition lorsque sont utilisés les poids équivalant à la valeur absolue de la distance mais n'a aucun effet sur ces fréquences lorsque ce sont les poids équivalant au carré de la distance qui sont utilisés. De façon générale, les transitions entre états non contigus ont plus de chances de survenir avec les poids équivalant à la valeur absolue de la distance qu'avec ceux équivalant au carré de la distance (ou ceux fondés sur une fonction qui croît plus rapidement avec la distance). Dans les quatre exemples du tableau 1, les deux seuls cas où il y a transition entre deux états non contigus sont le passage de l'état (ii) à l'état (iv) (poids équivalant à la valeur absolue de la distance, fonction économique z, probabilité de .08) et le passage de l'état (i) à l'état (iii) (poids équivalant à la valeur absolue de la distance, fonction économique z', probabilité de .02). Cependant, ce dernier cas est plus probable que le passage de l'état (i) à l'état (ii) (de fait, cette transition est impossible selon le tableau pertinent), ce qui est peut-être irréaliste.

Par ailleurs, la valeur des éléments diagonaux est généralement plus élevée lorsqu'on utilise les poids équivalant à la valeur absolue de la distance. Nous voyons d'après le tableau 1 que les poids équivalant à la valeur absolue de la distance produisent des éléments diagonaux qui sont égaux ou supérieurs aux éléments diagonaux obtenus à l'aide des poids équivalant au carré de la distance. Dans les quatre cas, la probabilité qu'un individu qui appartient à la classe (iv) y demeure est de 1.00, ce qui est probablement exagéré. Il ne faut pas en déduire qu'une personne qui a cessé de fumer ne recommencera jamais puisque les probabilités de transition sont différentes d'un groupe d'âge à l'autre.

La probabilité qu'un individu recommence à fumer (somme des éléments (4,2) et (4,3)) est nulle dans les quatre exemples et la probabilité qu'un individu cesse de fumer (somme des éléments (2,4) et (3,4)) est .08 ou .07. Lorsque nous avons tenté d'inverser cette relation (c'est-à-dire que la probabilité que l'on cesse de fumer soit inférieure à la probabilité que l'on recommence), nous avons provoqué une interruption du programme PL à certaines occasions car il n'existait pas de solution réalisable pour ces données à ces conditions. La difficulté venait non pas de la méthode PL mais de l'utilisation de données transversales et d'hypothèses inexactes à propos de ces données.

Il est utile d'examiner les probabilités de transition et d'analyser les conséquences d'une modification des paramètres, comme cela se fait pour le lissage des séries chronologiques. On peut analyser les avantages relatifs des diverses solutions et choisir la solution qui convient le mieux pour le modèle de microsimulation. La validité d'une série de probabilités de transition dépendra largement de la série de données et des hypothèses. Par exemple, l'hypothèse selon laquelle un individu ne peut passer directement de l'état (i) à l'état (iv) de la variable 4 est peut-être trop rigoureuse pour des tranches d'âge de cinq ans.

La figure 1 montre deux cycles d'évolution synthétiques simulés au moyen des distributions multidimensionnelles des quatre variables. En ce qui concerne la figure 1A, nous nous sommes servis des onze tableaux de probabilités de transition à huit dimensions, obtenus à l'aide de la fonction économique z et des poids équivalant au carré de la distance, pour imputer des états à un individu (que nous appellerons "Régis"). En ce qui concerne la figure 1B, nous nous sommes servis uniquement des distributions correspondant à chaque groupe d'âge, de sorte que l'état multidimensionnel d'un individu (que nous appellerons "Irenée") à l'âge t est indépendant de son état à l'âge t+1. Le cycle d'évolution de Régis est nettement plus régulier que celui d'Irenée. Régis escamote un état à une seule occasion tandis que chez Irenée, l'escamotage est fréquent. Dans le cas d'Irenée, nous relevons deux transitions invraisemblables en ce qui concerne l'usage du tabac; en effet, le graphique le classe, autour de 52 ans, parmi les personnes n'ayant jamais fumé alors qu'il a déjà fumé avant l'âge de 50 ans et le classe ensuite, à la tranche d'âge suivante, parmi les personnes qui ne fument plus. Son indice de masse corporelle fluctue de façon irréaliste, tout comme son taux de cholestérol. L'indice de masse corporelle de Régis évolue d'une manière tout à fait vraisemblable (hausse graduelle jusqu'à l'âge de 50 ans, puis diminution graduelle); même chose pour le taux de cholestérol. En ce qui a trait à l'usage du tabac, Régis suit une tendance normale, augmentant progressivement sa consommation de tabac jusqu'à l'âge de 50 ans, où il cesse de fumer. En revanche, le comportement d'Irenée est inexplicable en ce qui a trait à l'usage du tabac. Quant à la tension artérielle, elle suit une évolution normale dans les deux cas.

OBSERVATIONS RELATIVES AU CALCUL

Nous avons mentionné plus haut que les fréquences de transition avaient été calculées à l'aide d'une méthode de programmation linéaire. De fait, la question à laquelle nous nous intéressons dans cet article peut être assimilée à un domaine très particulier de la programmation linéaire, qui est la planification par réseaux, et à un sous-domaine de celle-ci, qui est le problème de transport. L'expression "problème de transport" vient de l'expression originale qui désignait une méthode permettant de trouver la solution la plus économique pour acheminer du matériel depuis les points d'approvisionnement jusqu'aux points de consommation (voir Hitchcock (voir par exemple Rogers (1980)). Lorsqu'on dispose de données longitudinales pour une population fermée, on connaît le nombre d'individus vivants et décédés pour n'importe quelle période donnée et on connaît aussi les fréquences de transition d'un état à l'autre; en revanche, les données transversales nous indiquent le nombre d'individus vivants mais non le nombre d'individus décédés ni les fréquences de transition. On peut montrer que le rajustement par le facteur C, du nombre d'individus vivants à l'âge t revient à soustraire de la population vivante d'âge t tous les individus qui mourront à l'âge t+1 mais en utilisant pour cela le taux de mortalité global au lieu du taux par état. Puisque le facteur de proportionnalité ne change rien aux résultats obtenus par les méthodes de programmation linéaire, cette opération de rajustement suppose l'utilisation d'un taux de mortalité indépendant de l'état.

C'est ce qu'on observe lorsque des données transversales sont rajustées; les probabilités de transition obtenues après ce rajustement peuvent être considérées comme des probabilités de transition de l'âge t à l'âge t+1 qui ne s'appliquent qu'aux individus encore vivants à l'âge t+1 mais cela, suivant l'hypothèse que tous les taux de mortalité par état sont égaux. L'utilisation de données transversales rajustées n'est pas très indiquée car si l'on définit plusieurs états pour une variable, c'est que l'on croit que le taux de mortalité varie selon l'état. Intuitivement, nous pouvons dire qu'un plus fort pourcentage d'individus à risque élevé mourront dans l'intervalle [t, t+1), ce qui fait qu'ils seront relativement moins nombreux à l'âge t+1. En supposant le même taux de mortalité pour tous les états, nous laissons croire que des individus à risque élevé sont passés à un état qui comporte moins de risques.

EXEMPLES

Les données dont nous nous servons pour illustrer la technique de lissage sont tirées de l'Enquête Santé Canada (ESC) de 1978-1979. Il s'agit d'une enquête à plan d'échantillonnage stratifié à plusieurs degrés, menée auprès de 31 668 individus. Pour en savoir plus sur l'ESC, le lecteur est prié de se reporter à Statistique Canada et Santé et Bien-être social Canada (1981). Pour chaque combinaison âge-sexe utilisée (12 groupes d'âge: 15-19, 20-24, 25-29, ..., 65-69, 70+), on a relevé des fréquences pour les variables et les classes suivantes:

Variable 1:

Indice de masse corporelle $(\frac{kg}{m^2})$

(i)	<20
(ii)	[20,25]
(iii)	(25,27]
(iv)	(27,30]
(v)	>30

Variable 3:

Tension artérielle diastolique (mmHG)

(i)	<90
(ii)	[90,105)
(iii)	≥105

Variable 2:

Taux de cholestérol sérique

(i)	<200
(ii)	(200,240]
(iii)	240

Variable 4:

$(\frac{mg}{dL})$

Usage du tabac

(i)	N'a jamais fumé
(ii)	1-20 cigarettes par jour
(iii)	Plus de 20 cigarettes par jour
(iv)	Ne fume plus

Les fréquences ont été calculées au moyen des poids de l'enquête. Les quatre variables ci-dessus sont des facteurs de risque qui peuvent servir à prédire les cas d'insuffisance coronaire. Les probabilités de transition calculées dans cet exemple doivent servir dans un modèle de microsimulation de la santé élaboré par Wolfson (1989); un sous-modèle construit par Wolfson et Birkett (1989) permet de simuler l'apparition et la progression de l'insuffisance coronaire.

Les transitions relatives à la variable 4 (usage du tabac) sont soumises par définition à certaines contraintes. La probabilité qu'un individu passe directement de l'état (i) à l'état (iv) est nulle (s'il est question d'une période relativement courte, où il n'y a le temps que pour une seule transition). La probabilité qu'un individu se retrouve dans la classe (i) après avoir été dans l'une ou l'autre des trois autres classes est nulle. De plus, il est raisonnable de supposer que la probabilité qu'un individu cesse de fumer est tout au plus égale à la probabilité qu'il recommence à fumer. Les éléments pertinents du tableau des probabilités de transition doivent donc satisfaire certaines équations ou inégalités. Par la méthode de programmation linéaire, il est possible de respecter ces relations en en faisant des contraintes.

Le tableau 1 donne un exemple d'application de la méthode de programmation linéaire pour une variable. En nous servant des distributions marginales observées pour la variable 4 (usage du tabac), nous avons calculé les fréquences et les probabilités de transition pour les hommes de 30 à 34 ans et de 35 à 39 ans. Les quatre états de la variable ont été classés selon l'ordre dans lequel ils sont susceptibles de se succéder dans la vie d'un les quantités $\{\frac{n_{t+1}}{n_t} u_{i_1 i_2}\}$ et $\{v_{j_1 j_2}\}$ comme les totaux marginaux $\{x_{i_1 i_2 ..}\}$ et $\{x_{..j_1 j_2}\}$ respectivement du tableau $\{x_{i_1 i_2 j_1 j_2}\}$ des fréquences de transition inconnues qui doivent être déterminées à l'aide des méthodes de programmation linéaire. Comme nous le verrons plus loin,les probabilités de transition qui découlent de ce calcul sont indépendantes de la valeur du facteur C.

La fréquence de transition $x_{i_1 i_2 j_1 j_2}$ représente le nombre d'individus (inconnu) qui sont passés de l'état (i_1, i_2) à l'âge t à l'état (j_1, j_2) à l'âge t+1. Le total général des fréquences de transition est $x_{....} = Cn_t = n$. La probabilité de transition $p_{i_1 i_2 j_1 j_2}$ est la probabilité qu'un individu se trouve dans l'état (j_1, j_2) à l'âge t+1 étant donné qu'il était dans l'état (i_1, i_2) à l'âge t:

$$p_{i_1 i_2 j_1 j_2} = \frac{x_{i_1 i_2 j_1 j_2}}{x_{i_1 i_2 ..}}.$$

(Le mot "probabilité" est utilisé sans trop de rigueur dans cet article et peut vouloir dire "proportion".) Notre but est d'obtenir des valeurs de $p_{i_1 i_2 j_1 j_2}$ (ou de $x_{i_1 i_2 j_1 j_2}$) qui soient acceptables pour produire des données microsimulées.

À l'aide des méthodes de programmation linéaire courantes, nous déterminons les valeurs $\{x_{i_1 i_2 j_1 j_2}\}$ de manière à minimiser une fonction économique qui équivaut à une somme pondérée des $x_{i_1 i_2 j_1 j_2}$ assujettie à trois types de contraintes: i) les fréquences doivent être non négatives; ii) les totaux marginaux des données multinomiales de départ doivent demeurer les mêmes et iii) les relations inhérentes aux variables doivent être respectées (p. ex.: le nombre d'individus passant de l'état "fumeur" à l'état "n'a jamais fumé" est nul).

Les poids de la fonction économique sont choisis de manière à favoriser la stabilité, c'est-à-dire à réduire au maximum le nombre de passages entre états non contigus (en supposant que la notion de "distance" entre états ait une signification). Si les indices d'état sont dans le bon ordre, on pourrait choisir comme poids $w_{i_1 i_2 j_1 j_2}$ de $x_{i_1 i_2 j_1 j_2}$ la distance entre l'état (i_1, i_2) à l'âge t et l'état (j_1, j_2) à l'âge t+1, par exemple $|i_1 - j_1| + |i_2 - j_2|$ ou $(i_1 - j_1)^2 + (i_2 - j_2)^2$.

Reste à savoir quelles variables utiliser: les fréquences de transition ou les probabilités de transition.

Autrement dit, la minimisation de

$$z = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} \, x_{i_1 i_2 j_1 j_2}$$

ne donne pas habituellement les mêmes résultats que la minimisation de

$$z' = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} p_{i_1 i_2 j_1 j_2} = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w'_{i_1 i_2 j_1 j_2} x_{i_1 i_2 j_1 j_2}$$

(où $w'_{i_1 i_2 j_1 j_2} = \frac{w_{i_1 i_2 j_1 j_2}}{x_{i_1 i_2 ..}}$).

Peu importe le facteur C utilisé, l'application d'une méthode de programmation linéaire à une série d'observations donnera toujours le même tableau de probabilités de transition (mais non le même tableau de fréquences de transition). Si $\{x_{i_1 i_2 j_1 j_2}\}$ sont les fréquences de transition obtenues avec la constante $C = C_1$, il est facile de montrer que les fréquences de transition obtenues avec la constante $C = C_2$ sont $\{\frac{C_2}{C_1} x_{i_1 i_2 j_1 j_2}\}$. Dans les deux cas, les tableaux de probabilités de transition sont les mêmes.

À ce stade-ci, il serait intéressant d'introduire une nouvelle variable (mortalité) pour laquelle il existe deux états: vivant et décédé. (Le second est un état absorbant; les personnes qui sont décédées demeurent "à jamais" dans le même état multidimensionnel et leur âge est défini comme le nombre d'années écoulées depuis la naissance.) Si l'on envisage la question sous l'angle inverse - c'est-à-dire que l'on imagine une table de mortalité à laquelle ont été ajoutées des variables qui représentent d'autres formes de transition que la mort - les fréquences de transition peuvent être assimilées à des éléments d'une table de mortalité multidimensionnelle

Par exemple, supposons que l'on se soit servi de données de référence transversales pour estimer à chaque âge la distribution d'une variable qui décrit l'usage du tabac chez un individu (trois états: "n'a jamais fumé", "fumeur" et "ne fume plus"). Si le modèle de microsimulation produit des données pour chaque âge considéré individuellement, les résultats simulés pourraient bien présenter des changements d'état anormalement fréquents de même que des cas de transition inconcevables (p. ex.: un individu qui passe de la catégorie "fumeur" à la catégorie "n'a jamais fumé").

Idéalement, un modèle de microsimulation utiliserait un tableau de probabilités de transition multidimensionnelles pour faire avancer un individu à travers les diverses tranches d'âge. Faute de données de référence multidimensionnelles, les analystes peuvent raccorder artificiellement différents fichiers de données (se servant des données relatives à un individu pour enrichir les données relatives à un autre individu ayant les mêmes caractéristiques) et peuvent en venir à supposer l'indépendance des variables distinctes.

Dans cet article, nous allons voir comment construire des tableaux de probabilités de transition multidimensionnelles à l'aide de données transversales multidimensionnelles dans le but de lisser le comportement longitudinal des individus simulés. Les exemples que nous présentons utilisent des données relatives à quatre variables pour lesquelles il existe respectivement 5, 3, 3 et 4 états, de sorte que nous aurons 180 fréquences à chaque groupe d'âge et 32,400 (180 x 180) probabilités de transition d'un groupe d'âge à l'autre. Comme il y a en tout 12 groupes d'âge, nous aurons donc 356,400 probabilités de transition. S'il existe des données multidimensionnelles pour deux groupes d'âge voisins, on peut construire un tableau de fréquences de transition multidimensionnelles (et le tableau de probabilités de transition correspondant) à l'aide de méthodes de programmation linéaire (PL). On fait en sorte que ces fréquences concordent avec les données de référence transversales multidimensionnelles et on impose aussi des conditions inhérentes à chacune des variables; cela, ajouté à la non-négativité des fréquences, constitue les contraintes du programme linéaire. La fonction économique du programme linéaire est choisie de manière que les transitions se fassent le plus possible entre états contigus (ce qui est raisonnable si relativement peu de temps sépare les deux groupes d'âge).

Il s'agit plus ici d'un problème de lissage que d'un problème d'estimation étant donné le très grand nombre de degrés de liberté dont nous disposons pour calculer les fréquences de transition et le nombre relativement restreint de totaux marginaux. Notre façon d'aborder le problème est analogue à celle utilisée pour le lissage des données chronologiques unidimensionnelles, pour lesquelles il existe de nombreux algorithmes de lissage; le choix de l'algorithme et des valeurs des paramètres se fait souvent de façon heuristique afin d'obtenir la qualité et le degré de lissage voulus. C'est dans cet esprit que nous proposons ici une méthode pour obtenir des microdonnées longitudinales convenablement lisses.

LA TECHNIQUE DE LISSAGE

Supposons que nous avons k variables d'intérêt ($k \geq 1$) pour lesquelles il existe des données multinomiales comme suit: pour chaque variable, on a défini un ensemble fini de tous les résultats possibles qui s'excluent mutuellement (états) et on a observé la fréquence de chaque combinaison d'états pour n_t individus d'âge t et n_{t+1} individus d'âge t+1. Si les données sont transversales, les deux groupes d'individus sont disjoints et n_{t+1} peut même être plus grand que n_t (ce qui ne peut être le cas dans une population fermée).

Le tableau des fréquences de transition de l'âge t à l'âge t+1 (et le tableau des probabilités de transition correspondant) est de dimension 2k. Pour simplifier la notation, nous supposons, sans perte de généralité, que k est égal à 2. Supposons maintenant que le nombre d'états pour la variable 1 est s_1 et le nombre d'états pour la variable 2 est s_2. Soit $u_{i_1 i_2}$ le nombre d'individus qui se trouvaient dans l'état bidimensionnel (i_1, i_2) à l'âge t et soit $v_{j_1 j_2}$ le nombre d'individus qui se trouvaient dans l'état (j_1, j_2) à l'âge t+1. (i_1 et j_1 désignent l'état pour la variable 1 et i_2 et j_2, l'état pour la variable 2; i_1 et $j_1 = 1, \ldots, s_1$; i_2 et $j_2 = 1, \ldots, s_2$). Alors, $n_t = u_{..}$ et $n_{t+1} = v_{..}$ (où un point signifie la sommation par rapport à l'indice inférieur indiqué). Normalement, $n_t \neq n_{t+1}$; cela est observé dans une population fermée à cause des décès et dans des données transversales parce que les deux groupes ne comprennent pas les mêmes personnes.

Supposons pour le moment qu'il ne se produit pas de décès entre l'âge t et l'âge t+1 et rajustons les fréquences observées (pour l'un ou l'autre des deux âges ou les deux) de manière que le nombre (n) d'individus soit le même pour chaque âge: multiplions les $u_{i_1 i_2}$ par une constante C et les $v_{j_1 j_2}$ par $C\frac{n_t}{n_{t+1}}$. Par exemple, si nous multiplions chaque fréquence observée à l'âge t par $C = \frac{n_{t+1}}{n_t}$, les fréquences observées à l'âge t+1 ne changent pas. Comme les deux séries de fréquences rajustées ont maintenant la même somme, nous pouvons considérer

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MÉTHODES DE LISSAGE POUR MICRODONNÉES LONGITUDINALES SIMULÉES

J.F. Gentleman et D.Robertson[1]

RÉSUMÉ

Les modèles de microsimulation permettent d'étudier le comportement de grandes populations dans le temps. À Statistique Canada, on a intégré des caractéristiques de santé et des facteurs de risque à un modèle démographique de la population active canadienne. Dans cet article, nous décrivons une méthode permettant de calculer des probabilités de transition multidimensionnelles utilisées dans la simulation de cycles de vie d'individus. Faute de données longitudinales, on doit se servir de données transversales pour calculer ces probabilités. Un modèle de microsimulation qui utilise des probabilités de transition adéquates produira des cycles de vie plus uniformes, plus vraisemblables. Pour qu'il y ait concordance avec les distributions transversales, des contraintes sont imposées aux probabilités. Ces contraintes peuvent être formulées comme celles du problème de transport dans la théorie de la planification par réseaux. Dans ce genre particulier de programmation linéaire, on choisit la fonction économique de manière à maintenir le degré et la fréquence des changements d'état dans le temps à un niveau raisonnable. À l'aide de données de l'Enquête Santé Canada, nous construisons des tableaux de probabilités de transition multidimensionnelles pour la consommation de tabac, la tension artérielle, le taux de cholestérol et l'indice de masse corporelle, qui sont tous reconnus comme d'importants facteurs de risque en ce qui a trait à l'insuffisance coronarienne.

MOTS CLÉS: Données longitudinales, microsimulation, simulation, lissage.

INTRODUCTION

Cet article met en évidence des techniques grâce auxquelles un modèle de microsimulation dynamique fondé sur des données transversales produit néanmoins des microdonnées longitudinales ajustées de manière réaliste. Un modèle de microsimulation est constitué d'une série d'algorithmes et d'un programme d'ordinateur destinés à la simulation de microdonnées. Les algorithmes reposent sur des sous-modèles probabilistes ou déterministes ou des distributions de données réelles. Le modèle de microsimulation produit un échantillon d'unités simulées qui représentent une population conceptuelle d'unités. Il peut s'agir, par exemple, de personnes, de ménages ou d'entreprises. Nous appellerons ces unités des "individus". L'échantillon d'individus sert à faire des inférences sur la population. Les modèles de microsimulation sont particulièrement utiles pour l'analyse prédictive par simulation. Afin de distinguer les données qui servent à la construction d'un modèle de microsimulation de celles produites par un tel modèle, nous appellerons les premières "données de référence" et les secondes "données simulées" ou "données d'échantillon". Orcutt, Merz et Quinke (1986) ont colligé toute une série d'articles utiles sur la microsimulation.

Un modèle de microsimulation dynamique permet de suivre l'évolution d'un échantillon d'individus dans le temps par la simulation de données multidimensionnelles (p. ex.: état matrimonial, statut professionnel, niveau de scolarité, consommation de produits manufacturés et état de santé) qui décrivent ces individus à chaque période de leur vie. De nombreuses enquêtes (par panel ou autres) peuvent produire des données de référence qui sont à la fois multidimensionnelles et longitudinales mais on ne voit pas toujours l'utilité, pourtant essentielle, de ces données pour un modèle de microsimulation. Selon Hoschka (1986, p. 49), l'absence de certaines variables et l'utilisation d'enquêtes transversales au lieu d'enquêtes par panel comptent parmi les facteurs qui amoindrissent le plus les données de référence des modèles de microsimulation. Nous savons que la collecte de données longitudinales est une opération qui s'étend nécessairement sur une longue période; or, il n'est pas toujours possible de savoir à l'avance sur quelles combinaisons de variables portera l'analyse. C'est pourquoi il faut envisager d'autres méthodes.

Supposons qu'un nombre fini de résultats (ou de classes, ou d'états) ont été définis pour chaque variable d'intérêt. À l'aide de microdonnées de référence longitudinales pour chaque âge, nous pouvons estimer la distribution d'une variable à un âge donné t de même que les probabilités de transition pour un individu qui passe d'un certain état à l'âge t à un autre état à l'âge t+1. Ces probabilités peuvent ensuite être utilisées dans le modèle de microsimulation pour suivre l'évolution de l'échantillon d'individus dans le temps.

Faute de données de référence longitudinales, les analystes se servent souvent de données transversales en considérant les données recueillies pour chaque groupe d'âge à un moment précis comme des données qui décrivent un groupe d'individus dans le temps. Normalement, on ne peut calculer des probabilités de transition à l'aide de données transversales (on ne peut non plus le faire avec des données longitudinales qui ne sont pas couplées d'une période à l'autre). Cependant, si un modèle de microsimulation ne tient pas compte du phénomène de transition et produit des données pour chaque groupe d'âge considéré de façon indépendante, les caractéristiques d'un individu simulé pourraient varier invraisemblablement d'une période à l'autre même si la distribution de l'échantillon concorde avec la distribution des données de référence pour chaque groupe d'âge.

[1] Division des études sociales et économiques, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6

ANNEXE

Mois	Date de déclarationau CDC				
	Janv. 87	Juil. 87	Janv. 88	Juil. 88	Janv. 89
Avant 1982	337	341	352	348	381
(Janv. 82) 1	50	50	48	49	54
2 2	64	66	68	67	68
3	58	59	59	59	62
4	59	58	57	56	58
5	59	61	61	64	68
6	71	71	72	74	76
7	81	84	84	85	88
8	92	94	95	94	96
9	109	105	106	107	108
10	104	106	106	106	111
11	118	123	121	125	125
12	134	135	137	137	139
13	170	174	172	174	180
14	149	153	151	153	161
15	200	203	207	212	216
16	212	214	216	222	229
17	211	217	219	222	225
18	257	259	260	259	263
19	229	234	238	239	245
20	243	244	248	251	249
21	259	264	264	266	272
22	245	249	257	261	261
23	277	278	277	277	277
24	313	318	315	318	321
25	343	343	354	358	362
26	368	373	381	380	389
27	383	385	384	397	403
28	414	424	432	434	441
29	454	464	469	470	474
30	434	447	451	456	463
31	484	490	495	503	513
32	500	507	516	517	528
33	506	513	516	525	534
34	555	567	574	577	582
35	530	538	551	551	559
36	560	568	577	589	597
37	630	650	657	678	701
38	611	623	634	647	655
39	728	744	757	780	799
40	751	766	784	817	838
41	755	770	796	815	834
42	773	799	819	853	879
43	889	906	936	965	984
44	954	976	1018	1057	1071
45	794	825	871	902	920
46	880	920	961	1018	1050
47	838	871	918	962	989
48	881	906	958	1014	1045
49	986	1059	1123	1197	1234
50	940	999	1072	1151	1191
51	966	1048	1135	1216	1263
52	990	1068	1145	1238	1278
53	1035	1123	1241	1326	1382
54	1034	1164	1276	1396	1462
55	1017	1176	1331	1437	1497
56	982	1185	1308	1425	1494
57	894	1205	1343	1495	1577
58	761	1292	1468	1602	1706
59	373	1097	1265	1391	1464
60	43	1167	1397	1548	1627
61	.	1228	1499	1707	1811
62	.	1164	1483	1694	1815
63	.	1172	1548	1709	1946
64	.	874	1542	1816	1927
65	.	523	1564	1864	2012
66	.	81	1532	1873	2036
67	.	.	1508	1915	2094
68	.	.	1456	1905	2085
69	.	.	1374	1950	2130
70	.	.	1117	1933	2146
71	.	.	499	1776	1974
72	.	.	58	1877	2132
73	.	.	.	1765	2073
74	.	.	.	1654	2064
75	.	.	.	1763	2317
76	.	.	.	1296	2026
77	.	.	.	840	2061
78	.	.	.	125	2167
79	.	.	.	.	1959
80	.	.	.	.	1982
81	.	.	.	.	1656
82	.	.	.	.	1352
83	.	.	.	.	759
84	.	.	.	.	122

Tableau A. Cinq séries de données fournies sur une base semestrielle par le CDC à compter de janvier 1987. Ces données correspondent aux chiffres mensuels des cas de SIDA diagnostiqués et comptabilisés à la date de déclaration.

pour quelque raison que ce soit, les périodes d'incubation seront allongées de manière artificielle et par conséquent, les temps de survie diminueront eux aussi de façon artificielle. L'inverse se produirait s'il s'avère que les diagnostics sont posés plus tôt que ne le prévoit la méthode des cinq stades de Walter Reed. Si les diagnostics ont été retardés par le passé, peut-être pour éviter des effets sociaux négatifs, et s'ils sont maintenant posés plus tôt, peut-être à cause de la promesse d'un temps de survie plus long à la suite de la découverte de remèdes non testés contre le SIDA, alors des effets trompeurs et contradictoires s'exercent sans doute sur la durée des périodes d'incubation et de survie. Il est possible que les temps de survie soient de fait diminués à cause de remèdes non testés et probablement coûteux, alors qu'ils peuvent sembler plus longs en raison d'un effet trompeur plus important du genre de ceux énoncés ci-dessus. Le besoin de réglementer ces remèdes ou cures miracles est un autre argument qui justifie la réalisation d'essais cliniques adéquats.

9. CONCLUSION

Si la science médicale ne réussit pas à découvrir des traitements, vaccins ou médicaments efficaces, l'humanité n'aura d'autre choix que de faire face au SIDA de la même façon qu'elle a réussi à survivre à d'autres épidémies par le passé, soit en acquérant une immunité naturelle. Au 16e siècle, les maladie infantiles amenées dans le Nouveau-Monde par les Espagnols ont eu des effets dévastateurs sur les populations autochtones des Amériques. Cependant, aujourd'hui, ces maladies n'ont pas plus de conséquences sur les descendants de ces population qu'elles en ont pour les descendants des Espagnols. Il a fallu plusieurs générations avant qu'une immunité naturelle ne puisse être acquise; six générations selon les estimations de W. H. McNeill (1975). Dans la mesure où le nombre des cas d'infections à VIH sont assez rares chez les enfants, il se peut qu'un plus grand nombre de générations soit nécessaire pour acquérir une immunité contre le SIDA. En attendant, le principal espoir dans la lutte à court terme contre l'épidémie repose sur l'éducation.

BIBLIOGRAPHIE

Brodt, H.R., E.B. Helm, A. Joetten, L. Bergmann, A. Kluver et W. Stille (1986), *Spontanverlauf de LAV HTLV-III-Infektion: Verlaufsbeobachtungen bei Personen aus AIDS-Risikogruppen*; Deutsche Medizinische Wochenscrift, Stuttgart, Vol. iii, pp. 1175-1180.

Chernoff, H. et S. Zacks (1964), "Estimating the current mean of a normal distribution which is subject to changes in time", *Annals of Mathematical Statistics* 35, 999-1018.

Cowell, M.J. et W.H. Hoskins (1987), AIDS, HIV Mortality and Life Insurance, parties 1 et 2, Society of Actuaries, distribué en tant que rapport spécial.

Duong, Q.P. et I.B. MacNeill (1987), Selection and estimation of growth models with application to forecasting AIDS. *Department of Statistical and Actuarial Sciences, Technical Report TR-87-09*. London, Canada: The University of Western Ontario.

Healy, M.J.R. et H.E. Tillet (1988), Short-term extrapolation of the AIDS epidemic, *Journal of the Royal Statistical Society, Series A*, 50-61.

Jandhyala, V.K. et I.B. MacNeill (1989), Detection of parameter changes at unknown times in linear regression models, *(à paraître)*.

Jandhyala, V.K. et I.B. MacNeill (1989), Change detection methodology for modelling the incidence of AIDS, *Department of Statistical and Actuarial Sciences, Technical Report TR-89-01*. London, Canada: The University of Western Ontario.

Johnson, A.M. (1988), Social and behavioural aspects of the HIV epidemic - A review, *Journal of the Royal Statistical Society, Series A*, 151, 99-114.

Kalbfleisch, J.D. et J.F. Lawless (1988), Inference based on retrospective ascertainment. An analysis of the data on transfusion related AIDS, *Technical Report STAT-88-02*, Department of Statistics and Actuarial Science, University of Waterloo.

MacNeill I.B. (1978), Properties of sequences of partial sums of polynomial regression residuals with application to tests for change in regression at unknown times, *Annals of Statistics* 6, 422-433.

MacNeill, I.B. (1989), The reporting-delay function, *Department of Statistical and Actuarial Sciences, Technical Report TR-89-04*, London, Canada: The University of Western Ontario.

Martin, J.L. (1987), The impact of AIDS on gay male sexual behaviour patterns in New York City, *American Journal of Public Health*, 77, 578-581.

McKusic, M., W. Horstman et J.J. Coates (1985), AIDS and sexual behaviour reported by gay men in San Francisco, *American Journal of Public Health*, 75, 493-496.

McNeill, W.H. (1976), Le temps de la peste, Doubleday: New York.

Morgan, W. Meade et James W. Curran (1986), Acquired Immunodeficiency Syndrome: current and future trends, *Public Health Report 101* 5, 459-465.

Panjer, H.H. (1987), AIDS: Survival analysis of persons testing HIV +, Working Paper series in Actuarial Science ACTSC87-14, Waterloo, Canada: The University of Waterloo.

"Public Health Service Plan for the Prevention and Control of AIDS and the AIDS virus", Report of the Coolfont Planning Conference, U.S. Public Health Service, Washington, 1986, p.1

Winkelstein, W., M. Samuel, N.S. Padrian et J.A. Whiley (1987), Select sexual practices of San Francisco heterosexual men and risk of infection by the human immunodeficiency virus, *Journal of American Medical Association*, 257, 1370-1471.

La caractéristique la plus importante des estimations du nombre des cas d'infection par le VIH est qu'après avoir atteint un sommet en 1986, il se soit mis à décliner brusquement par la suite. Plusieurs raisons peuvent expliquer cette chute marquée. Tout d'abord, l'éducation au sein des groupes à risque élevé a eu pour effet de modifier le comportement des membres de ces groupes. Toutefois, en raison de la longue durée de la période d'incubation moyenne, l'éducation n'a probablement eu qu'un effet négligeable sur la configuration de la courbe de la fonction N(t) au cours de la période antérieure à 1985-1986. Une autre explication, plus plausible, de la croissance rapide et du déclin qui a suivi est l'effet de saturation survenu parmi ceux qui étaient le plus à risque dans la période qui a précédé immédiatement le moment où la maladie est devenue connue de tous. En d'autres termes, il s'est produit au début une croissance de type exponentielle, mais une fois que l'infection a gagné une large part du groupe concerné, il n'y avait plus beaucoup de place pour le maintien d'une telle tendance.

Les effets du programme d'éducation ainsi que de la large diffusion d'informations au sujet du SIDA au début de la décennie se feront sans doute davantage sentir au cours de la prochaine décennie. Si ces programmes se révèlent efficaces, alors le scénario C est le plus probable; dans le cas contraire, c'est le scénario A qui est le plus vraisemblable.

Statistiques à l'appui, Johnson (1988), McKusick et coll. (1985), Martin (1987), Winkelstein et coll. (1987) ainsi que d'autres, ont démontré que l'éducation a entraîné un changement de comportement notable chez la population male homosexuelle des Etats-Unis. Par conséquent, compte tenu de ces faits, le scénario C devient le plus probable. Cependant, on connaît moins bien l'effet de l'éducation sur les utilisateurs de drogues.

Quoi qu'il en soit, même si le nombre de nouveaux cas d'infection devait cesser complètement d'augmenter, le SIDA demeurerait une épidémie importante tout au long des années 1990 simplement parce que le nombre de cas latents demeure relativement élevé.

Il est plus encourageant de se rendre compte que les effets de la montée rapide et du déclin marqué du taux de croissance de la population infectée par le VIH aux Etats-Unis, qui ont eu lieu avant la fin des années 1980, s'estomperont au cours des années 1990 et auront nettement diminué au tournant du siècle.

Dans l'analyse qui précède, nous avons posé comme hypothèse qu'il n'y aurait pas de découverte médicale majeure sous forme de traitements ou de vaccins contre le SIDA. Nous avons également supposé l'absence de découverte de médicaments ou de thérapies qui contribueraient à allonger la période d'incubation moyenne de l'infection à VIH et (ou) le temps de survie des victimes du SIDA. Evidemment, de tels traitements, vaccins, médicaments ou thérapies aurait un effet déterminant sur l'évolution de l'épidémie.

L'analyse repose aussi sur la notion selon laquelle la distribution des périodes d'incubation est stationnaire. Quel serait l'effet d'un allongement de la période d'incubation moyenne, attribuable par exemple à un traitement thérapeutique quelconque, à un moment donné dans l'avenir? La figure 12 présente l'illustration graphique de D(t) selon les scénarios A, B et C, où la période d'incubation moyenne est allongée de 10 à 15 années à compter de 1990. Les principaux effets d'un tel allongement sont une diminution de l'incidence de l'épidémie dans les années 1990, mais aussi une hausse du nombre de cas au début des années 2000.

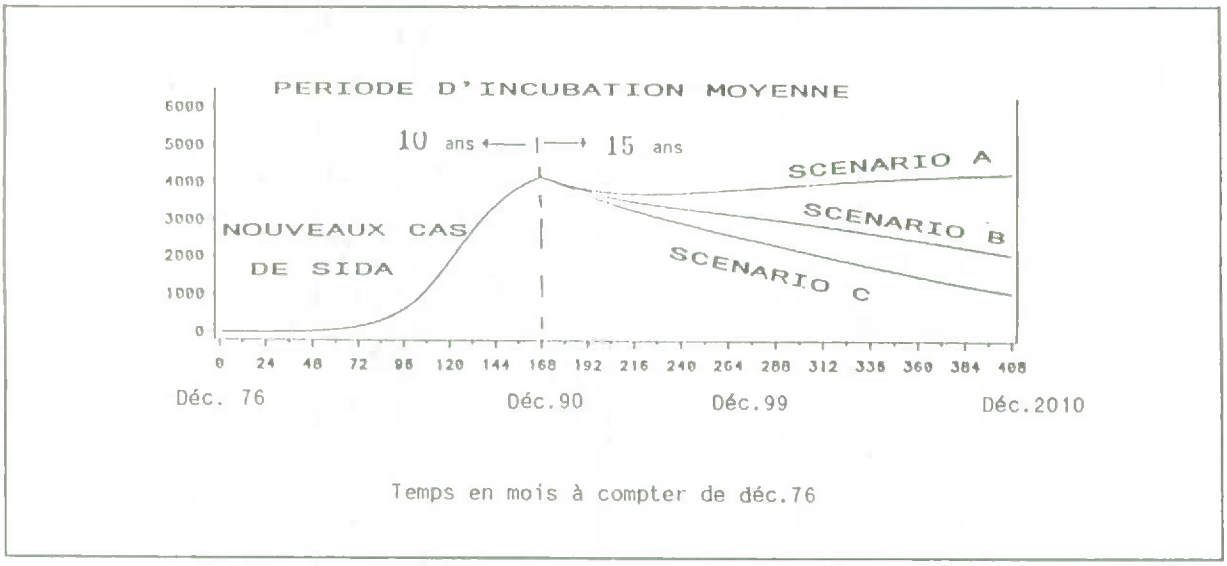


Figure 12. Nouveaux cas de SIDA selon trois scénarios concernant l'avenir (voir figure 10), avec augmentation de 10 à 15 années de la période d'incubation moyenne à compter de 1990.

Il est à remarquer que l'intervalle entre le moment de l'infection par le VIH et le moment du décès causé par le SIDA est réparti en deux périodes selon le moment du diagnostic du SIDA. La première de ces périodes correspond à la période d'incubation et la seconde au temps de survie. Si les diagnostics de SIDA sont retardés

Selon ces calculs, le nombre des cas latents de SIDA aux Etats-Unis se situerait aux environs de 600,000, un nombre sans doute considérable, mais beaucoup moins important que l'estimation du bureau du Surgeon General (1989).

Il est à noter qu'aux fins de la positivité de la série N(t), nous devons fixer une limite supérieure à la durée moyenne de la période d'incubation pour un modèle de distribution donné. Dans les formules que nous avons utilisées, afin que les distributions des périodes d'incubation coïncident sensiblement avec l'évolution que semble suivre le nombre de cas de SIDA diagnostiqués et compte tenu de l'équation (6) et de la positivité de (N(+), la période d'incubation moyenne ne pouvait guère dépasser 10 années.

7. PRÉVISION DU NOMBRE DES CAS DE SIDA

Que nous réserve l'avenir pour ce qui est du nombre de cas de SIDA aux Etats-Unis? Plusieurs scénarios peuvent être étudiés par extrapolation de N(t), la fonction du taux d'infection par le VIH; la figure 10 contient trois extrapolations différentes, chacune correspondant à une période moyenne d'incubation de 10 années. L'équation (6) peut ensuite être utilisée pour estimer les valeurs futures de la série des cas de SIDA diagnostiqués. L'extrapolation A est compatible avec le maintien de la croissance logistique de D(t), la série des cas de SIDA diagnostiqués. Les extrapolations B et C se rapprochent davantage de la dynamique interne apparente de la série N(t), et laissent présager une diminution du nombre des cas de SIDA diagnostiqués à compter de l'année 1990. Les prévisions correspondantes pour D(t), fondées sur les trois scénarios associés à N(t), sont présentées à la figure 11.

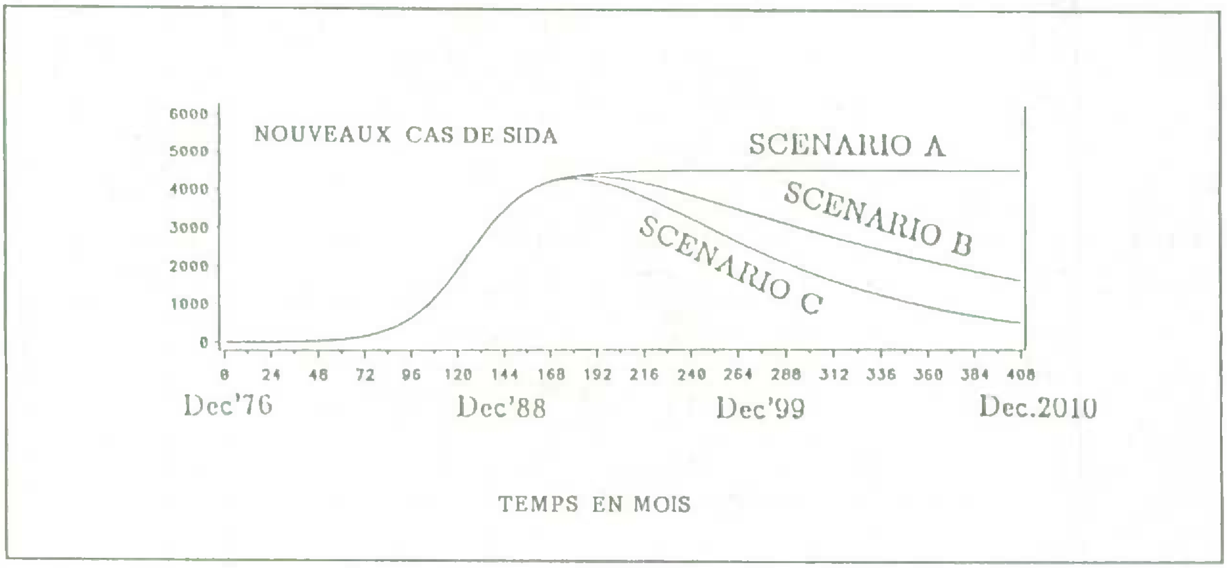


Figure 11. Nouveaux cas de SIDA selon trois scénarios concernant l'avenir (voir aussi la figure 10).

8. ANALYSE

Les estimations de la taille de la population américaine infectée par le VIH, résultats des calculs que nous venons de décrire, sont plus petites que celles données dans le rapport Coolfont (1986) et dans les rapports plus récents publiés par le bureau du Surgeon General.

Plusieurs facteurs influent de manière significative sur la taille des estimations fournies dans ce document. Premièrement, plus la période d'incubation est longue, plus grande est l'estimation de la taille de la population infectée par le VIH. Cela est attribuable au fait que le nombre de cas de SIDA observés jusqu'à maintenant représente une proportion de la population infectée qui varie en raison inverse de la durée de la période d'incubation de ces mêmes cas.

Deuxièmement, nous posons comme hypothèse que le pourcentage de la population infectée qui deviendra éventuellement atteinte du SIDA est de 100%. Si cette hypothèse est fausse et si le pourcentage est 100 p %, où $0 < p < 1$, les estimations de la taille de la population infectée doivent être augmentées par le facteur p^{-1}. Aucun ajustement est nécessaire si on ne s'intéresse qu'aux cas latents de SIDA.

Troisièmement, l'information concernant la sous-déclaration est hautement spéculative. Cependant, si la fraction de l'ensemble des cas déclarés est représentée par f, alors les estimations de la taille de la population infectée par le VIH doivent être augmentées par le facteur f^{-1}. Il est probable que l'effet de la sous-déclaration ne dépasse pas deux erreurs types de l'asymptote logistique.

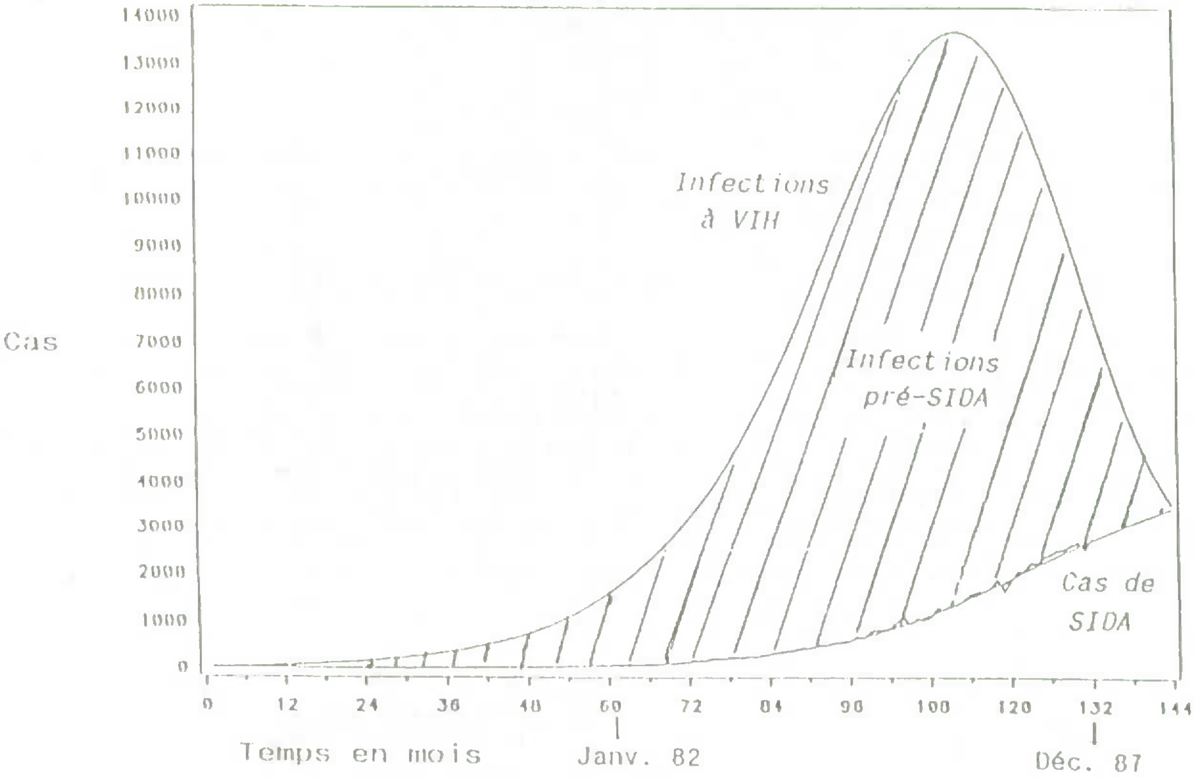


Figure 9. Taux de nouveaux cas d'infection à VIH, N(t), et taux des cas de SIDA diagnostiqués, D(t)

Périodes d'incubation moyennes	Infections à VIH pré-SIDA			
	1985		1988	
	Logistique	Logistique $+2\sigma$	Logistique	Logistique $+2\sigma$
8 années	232,500	466,800	443,300	498,400
9 années	278,800	558,600	515,100	567,500
10 années	329,000	657,700	590,600	638,000

Tableau 3. Estimations de la taille de la population au stade pré-SIDA aux États-Unis

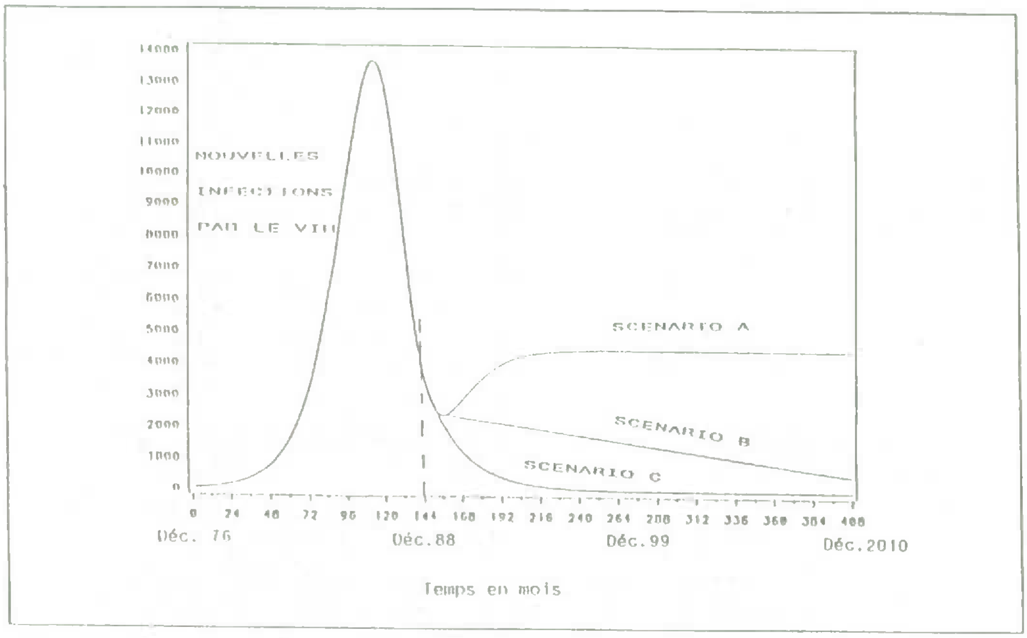


Figure 10. Nouvelles infections par le VIH selon trois scénarios concernant l'avenir

Le dernier stade est la mort.

Le tableau 1, qui présente certains des principaux résultats de l'étude, fait état du nombre de patients observés selon le stade et selon la durée de la période d'observation.

Périodes d'observation	Stade 1a (A risque)	Stade 1b (VIH+)	Stade 2a (SLA)	Stade 2b (Para-SIDA)	Stade 3 (SIDA)	Tous les stades
3-6 months	10	9	21	8	6	54
6-12 months	14	18	51	29	9	121
12-24 months	21	20	29	20	7	97
24-36 months	3	5	19	7	1[1]	35
All Periods	48	52	120	64	23	307

Tableau 1. Etude de Francfort, données du tableau 5 "Nombre de patients observés selon le stade et la période d'observation"

n	1	2	3	4	5	6	7	8	9	10	11	12
$P(j, j+n)$	0.016	0.065	0.107	0.125	0.124	0.113	0.096	0.080	0.063	0.050	0.039	0.030

Tableau 2. Proportion des personnes nouvellement infectées qui contracteront le SIDA ou en mourront n années plus tard

Ces données ont été utilisées par Cowell et Hoskins (1987) et par Panjer (1987) pour estimer les taux de progression des infections à VIH et des taux de mortalité due au SIDA. Nous avons utilisé le modèle mis au point par Panjer pour calculer les estimations de P(j,j,+n) présentées dans le tableau 2. Ce modèle suppose une distribution des périodes d'incubation qui s'approche de la loi gamma avec une moyenne de 6.3 années. Cowell et Hoskins, ayant analysé les mêmes données avec un modèle différent, ont obtenu une moyenne de près de deux années de plus. Kalbfleish et Lawless (1988) ont estimé que la période d'incubation médiane est d'environ 10 années. Nous nous sommes ensuite servis de la loi gamma en faisant varier la valeur des paramètres pour obtenir différents modèles de distribution des périodes d'incubation. Nous reviendrons plus loin sur les limites supérieures des périodes d'incubation moyennes que nous situons aux environs de dix (10) années.

6. ESTIMATIONS DE LA TAILLE DE LA POPULATION INFECTÉE PAR LE VIH AUX ÉTATS-UNIS

La figure 8 illustre les données mensuelles sur les cas de SIDA diagnostiqués aux Etats-Unis, lesquelles ont été ajustées pour tenir compte des déclarations tardives. Le modèle logistique après ajustement et extrapolation jusqu'à la fin de 1988 est utilisé pour estimer la taille de la population infectée par le VIH aux Etats-Unis. Ces estimations sont obtenues à la suite de l'application de la loi gamma à la distribution des périodes d'incubation et de l'utilisation des divers ensembles de paramètres définis dans les études citées à la section 5. L'équation (6) est résolue numériquement et produit les estimations présentées au tableau 3. La figure 9 illustre graphiquement $D(t)$, le taux des nouveaux cas de SIDA diagnostiqués (après ajustement du modèle logistique), et $N(t)$, le taux de nouveaux cas d'infection. Pour cette caractéristique, les paramètres de l'ajustement logistique sont $D(0) = 75.92$, $M = 4476.86 \pm 255.33$ et $k = 0.00001412$. Pour les données de la figure 9, nous avons supposé que la distribution des périodes d'incubation suit la loi gamma avec $\alpha = 2$ et $\beta = 5$ pour une période d'incubation moyenne de 10 années. Les calculs sont répétés après augmentation de l'asymptote de la courbe logistique de deux erreurs types.

La caractéristique la plus frappante du graphique de $N(t)$, la fonction du taux d'infection, est la hausse rapide du nombre des nouveaux cas d'infection jusqu'en 1985-1986 et le déclin tout aussi marqué des années subséquentes. Comme nous l'avons souligné, les calculs ont été répétés en fonction de diverses autres distributions possibles des périodes d'incubation et d'asymptotes plus grandes du modèle logistique. Dans tous les cas, la fonction du taux d'infection croît rapidement avant 1985-1986 et régresse nettement après 1986. Les autres facteurs étant constants, des périodes d'incubation moyennes plus longues ont pour effet d'augmenter la taille estimée de la population infectée par le VIH; cela est illustré dans le tableau 3.

canadiennes. Grâce à ce critère, il est possible de comparer des modèles non imbriqués et de faire des inférences au sujet de la sélection d'un modèle. Dans le cas des données canadiennes, nous avons choisi le modèle logistique, lequel peut servir à prédire l'incidence du SIDA; d'après des applications antérieures de la méthode, il semble que les prévisions obtenues soient assez précises, du moins pour ce qui est d'un proche avenir. A ce stade de l'évolution de l'épidémie, il est peu probable que les modèles réalistes non empiriques soient vraiment utiles à la prévision à court terme en raison du grand nombre de paramètres qu'ils comportent. Par conséquent, nous nous servirons uniquement du modèle logistique pour le lissage et pour les prévisions à court terme du nombre de cas de SIDA diagnostiqués.

La fonction logistique es définie de la manière suivante:

$$D(t) = \frac{MD(0)}{D(0) + (M - D(0))\exp(-mkt)} \quad ,$$

où $D(0)$ représente la taille de l'épidémie lorsque $t = 0$, M la taille maximale (taux) de l'épidémie, k le coefficient de la pente et $D(t)$ le taux de cas diagnostiqués au moment t. Cette fonction a fait l'objet d'un ajustement par régression non linéaire des moindres carrés au nombre de cas de SIDA diagnostiqués tels que comptabilisés en date de janvier 1989 par les Centers for Disease Control (CDC); décembre 1981 est considéré comme le temps $t = 0$. Les données pour 1988 n'ont pas été utilisées dans l'ajustement en raison du problème grave de déclaration tardive qui caractérise la déclaration des cas de SIDA. Cependant, les données antérieures à 1988 ont été corrigées à la hausse à l'aide de la méthode d'ajustement pour tenir compte de la déclaration tardive qui est décrite à la section 2. Le modèle logistique est utilisé ici uniquement pour les besoins de la prévision à très court terme, c'est-à-dire une période de douze (12) mois allant jusqu'en décembre 1988.

La figure 8 montre le graphique des données corrigées qui sont superposées sur la courbe du modèle logistique ajusté.

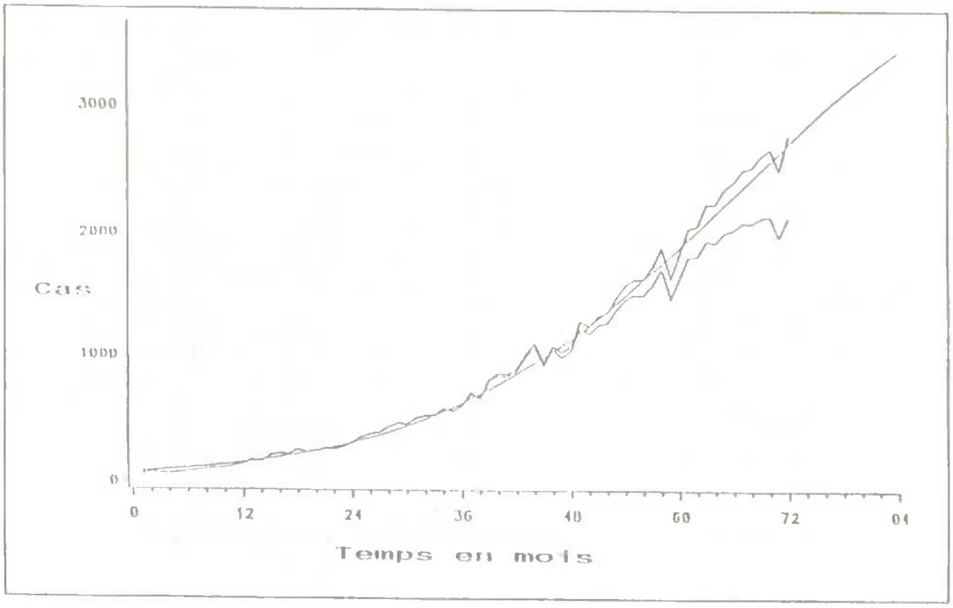


Figure 8. Série des cas de SIDA diagnostiqués (après ajustement), janvier 1982 - décembre 1987, tels que comptabilisés en date de janvier 1989, et courbe ajustée du modèle logistique avec extrapolation jusqu'en décembre 1988.

5. PÉRIODES D'INCUBATION DES INFECTION AU VIH

Une étude longitudinale a été menée par Brodt et coll. (1986) à l'université de Francfort auprès de sujets à risque pour ce qui est du SIDA dans le but de déterminer la durée de la progression de la maladie aux divers stades. Les auteurs de l'étude ont utilisé les cinq stades de la méthode de Walter Reed pour suivre l'évolution de l'état de personne saine à l'état de patient atteint du SIDA. Les cinq stades sont:

1a (A risque): Personnes saines à risque en ce qui concerne l'infection à VIH, mais dont le test est négatif.
1b VIH +: Personnes asymptomatiques ayant obtenu un test VIH positif.

2a SLA: Patients qui présentent une infection à VIH et un syndrome lymphadénophatique (SLA) de même qu'une déficience de l'immunité à médiation cellulaire de gravité moyenne.

2b Para-SIDA: Patients qui présentent une infection à VIH et un SLA de même qu'une grave déficience de l'immunité à médiation cellulaire (para-SIDA, tel que défini par le CDC).

3 SIDA: Patients atteints du SIDA tel que défini par le CDC.

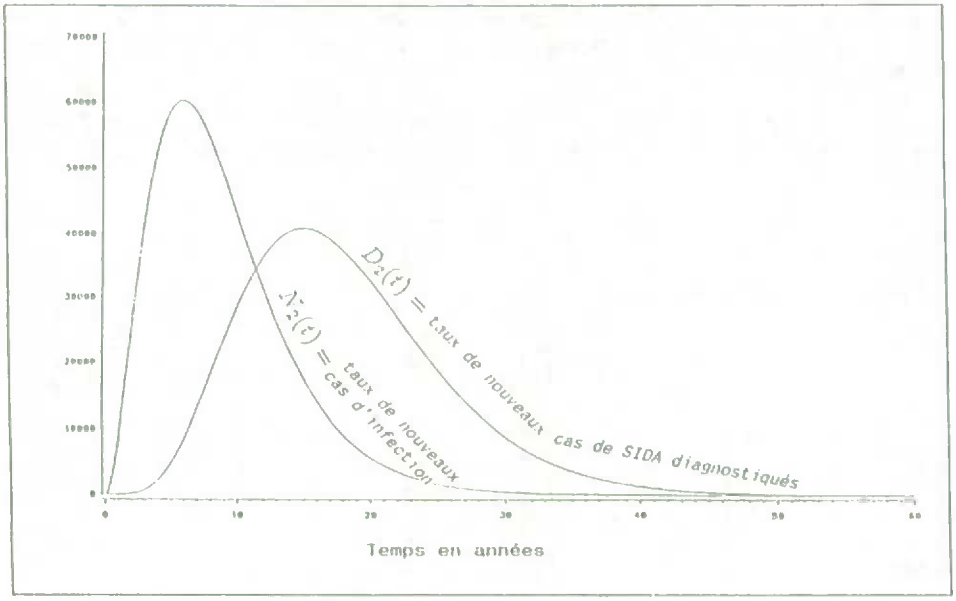


Figure 7. Taux de nouveaux cas de SIDA diagnostiqués et de nouveaux cas d'infection à VIH; modèle 2.

La solution par la transformation de Laplace donne

$$N_2(t) = \frac{K}{\Gamma(\alpha - a)\beta^{\alpha-a}} t^{\alpha-a-1}\exp(-t/\beta) \quad .$$

La figure 7 illustre $D_2(t)$ et $N_2(t)$ avec $\alpha = 6$, $a = 3$, $\beta = 3$, et $K = 699{,}840$. La période d'incubation moyenne que suppose ce choix de paramètres est de 9 années.

Les paramètres de chacun des exemples ci-dessus ont été choisis de façon à produire un total d'environ 80,000 cas de SIDA dans les dix (10) premières années de l'épidémie.

Dans l'éventualité de l'absence de solutions analytiques à l'équation intégrale (6), on pourrait recourir à des techniques numériques pour trouver des solutions. Ces méthodes, fondées sur les équations (5), peuvent être testées à l'égard des solutions exactes représentées aux figures 6 et 7. C'est ce qui a été fait et nous avons constaté qu'il était possible d'atteindre les niveaux de précision déterminés à l'avance.

4. NOMBRE DE CAS DE SIDA DIAGNOSTIQUÉS

Aux Etats-Unis, les premiers cas de SIDA ont été diagnostiqués en 1978; il est possible qu'il y ait eu des cas antérieurs, mais ces derniers n'ont pas été diagnostiqués. La taille de l'épidémie s'est ensuite mise à croître de façon exponentielle pendant plusieurs années. Cette forte croissance exponentielle du début a entraîné la prédiction d'une calamité aussi grande que la peste noire qui a décimé la population d'Europe pendant les années 1300. Ces prévisions ont été faites en fonction du maintien des tendances observées à ce moment.

Toutefois, le nombre des cas de SIDA diagnostiqués semble avoir cessé de croître de façon exponentielle au début de l'année 1984. Ce revirement a été observé par Duong et MacNeill (1987) dans le cas des données canadiennes, et Jandhyala et MacNeill (1988), après avoir analysé les données américaines, ont estimé que les paramètres du système ont commencé à changer au début de l'année 1984. La méthodologie utilisée pour vérifier l'hypothèse du changement des paramètres à un moment inconnu est expliquée par Chernoff et Zacks (1964), MacNeill (1978) et Jandhyala et MacNeill (1986). Ayant déterminé que l'hypothèse de la croissance exponentielle ne pouvait plus être soutenue, Duong et MacNeill (1986) ont utilisé le critère d'information d'Akaike pour choisir parmi une série de modèles de croissance celui jugé le plus approprié aux données

Ce choix de la fonction des taux des cas diagnostiqués suppose une croissance qui part de zéro au temps $t = 0$ et qui se rend jusqu'à l'asymptote K, le taux de croissance étant contrôlé par c. Qualitativement, une telle croissance est plausible d'un point de vue épidémiologique pour la tranche de temps à laquelle nous nous intéressons présentement.

Comme cas réel avec des périodes d'incubation distribuées selon la loi gamma, prenons $\alpha = 2$ et $\beta = 4$, ce qui suppose un période d'incubation moyenne de 8 années. La transformation de Laplace de $D_1(t)$ et de $P_1(t)$ s'écrit:

$$\mathcal{L}(D_1) = 2K \sqrt{c}\, p^{-\frac{1}{2}} K_1(2\sqrt{c}\, p^{\frac{1}{2}}),$$

où $K_1(.)$ est une fonction de Bessel modifiée, et

$$\mathcal{L}(P_1) = \frac{1}{16}(p + \frac{1}{4})^{-2} .$$

Alors,

$$\mathcal{L}(N_1) = 32K \sqrt{c}\, p^{-\frac{1}{2}} K_1(2 \sqrt{c}\, p^{\frac{1}{2}})(p + \frac{1}{4})^2 ,$$

dont l'inversion donne

$$N_1(t) = K \exp\{-c/t\}(1 + 8ct^{-2} -32ct^{-3} + 16c^2t^{-4}) .$$

La figure 6 illustre $D_1(t)$ et $N_1(t)$ avec $K = 36,000$ et $c = 9.5$.

Comme deuxième exemple, prenons

$$D_2(t) = \frac{K}{\Gamma(\alpha)\ \beta^\alpha} t^{\alpha-1}\exp\{-t/\beta\} \qquad t>0$$

et

$$P_2(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1}\exp\{-t/\beta\} , \qquad t>0.$$

Ce choix de la fonction des taux de SIDA diagnostiqués suppose une croissance qui part de zéro à $t = 0$, qui atteint un maximum et qui se met ensuite à décliner en s'approchant asymptotiquement de zéro; les taux de croissance est de déclin sont déterminés par α et β. Là aussi, une telle croissance est qualitativement plausible pour certaines séries épidémiologiques, mais il est peut-être prématuré de prédire le moment du revirement du nombre de cas de SIDA diagnostiqués; nous traitons de cette question plus loin, soit la section 7.

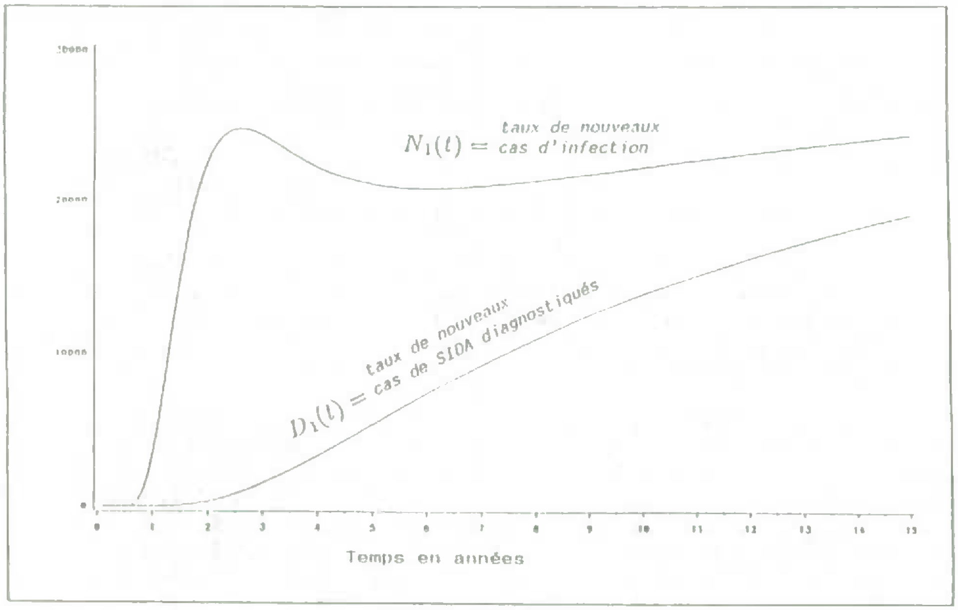


Figure 6. Taux de nouveaux cas de SIDA diagnostiqués et de nouveaux cas d'infection à VIH; modèle 1.

3. RELATION ENTRE LE NOMBRE DE CAS DIAGNOSTIQUÉS ET DE PERSONNES INFECTÉES ET LA PÉRIODE D'INCUBATION

Dans cette section, nous cherchons à construire des équations reliant le nombre de cas de SIDA diagnostiqués par unité de temps à l'intérieur d'une juridiction donnée et le nombre de personnes infectées par le VIH au cours d'unités de temps antérieures. Pour les besoins de l'analyse, nous utilisons l'année comme unité de temps. Supposons que $D(k)$ représente le nombre de cas de SIDA diagnostiqués au cours de l'année k. Supposons en outre que $I(j,k)$ représente le nombre de personnes infectées au cours de l'année j et pour lesquelles un diagnostic de SIDA a été posé au cours de l'année $k(k \geq j)$; par conséquent

$$D(k) = \sum_{j='76}^{k} I(j,k) .$$

Nous supposons ici que les première infections au VIH ont été contractées en 1976 ou plus tard. Aussi, si $P(j,k)$ représente la proportion du total des personnes infectées au cours de l'année j pour lesquelles un diagnostic de SIDA est posé subséquemment au cours de l'année k et si $N(j)$ représente le nombre total de personnes infectées au cours de l'année j, alors

$$P(j,k) = \frac{I(j,k)}{N(j)} .$$

Soit $T(1)$ le nombre de personnes infectées jusqu'à l'année l; il s'agit des valeurs pour lesquelles on semble connaître si peu de choses et pour lesquelles nous pouvons produire des estimations à l'aide du système d'équations:

$$T(1) = \sum_{j='76}^{k} N(k) \qquad 1 = '76,'77, \ldots,$$

$$D(k) = \sum_{j='76}^{k} N(j)\ P(j,k) \quad k = '76,'77, \ldots . \tag{5}$$

Comme nous l'avons vu à la section 2, nous disposons maintenant d'une information substantielle au sujet de $D(k)$, et d'autres estimations peuvent être obtenues grâce à la prévision; nous reviendrons plus en détail sur la série chronologique en question à la section 4. Une des premières études ayant fourni des données au sujet de $P(j,k)$ vient des travaux cliniques Brodt et coll. (1986) et de l'analyse de ces données par Cowell et Hoskins (1987) et Panjer (1987). On trouve d'autres estimations des périodes d'incubation dans des études plus récentes dont font état Kalbfleisch et Lawless (1988). La section 5 présente une analyse de l'estimation de $P(j,k)$. À la section 6, la méthodologie exposée à la présente section est appliquée à la série des cas diagnostiqués et à la distribution des périodes d'incubation pour obtenir des estimations de la taille de la population infectée par le VIH aux Etats-Unis.

Nous pouvons d'ores et déjà souligner que l'équation (5) est l'analogue discontinu de l'équation intégrale suivante

$$D(t) = \int_0^t P(t - s)N(s)ds \quad . \tag{6}$$

Par conséquent, si $D(.)$ et $P(.)$ sont connus, nous pouvons obtenir $N(.)$ en résolvant (6). Cela nous offre un outil très utile pour l'étude des relations entre modèles plausibles de la fonction des taux des cas diagnostiqués $D(.)$, de la fonction des taux d'infection $N(.)$ et de la distribution des périodes d'incubation $P(.)$.

Si (f) représente la transformation de Laplace de la fonction $f(.)$, alors l'équation (6) donne

$$\mathcal{L}(D) = \mathcal{L}(P)\mathcal{L}(N) .$$

Si $D(.)$ et $P(.)$ sont connus et si leurs transformations peuvent être calculées par analyse, alors

$$\mathcal{L}(N) = \mathcal{L}(D)/\mathcal{L}(P) .$$

La distribution des taux d'infection $N(.)$ peut ensuite être obtenue par l'inversion de sa transformation. A titre d'exemple, prenons les fonctions

$$D_1(t) = K \exp(-c/t), \quad t>0 ,$$

et

$$P_1(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1}\exp(-t/\beta), \quad t>0 .$$

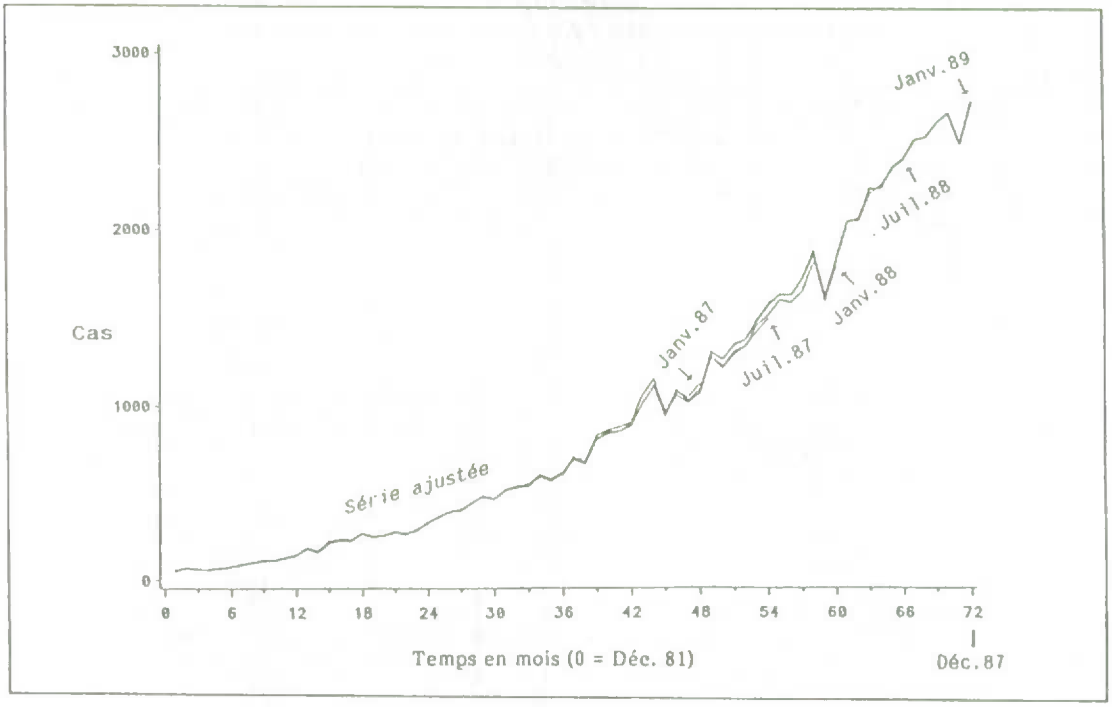


Figure 4. Série des cas de SIDA diagnostiqués aux E.-U., tels que comptabilisés en janvier 1989, ajustée pour tenir compte des déclarations tardives, janvier 1982-décembre 1987.

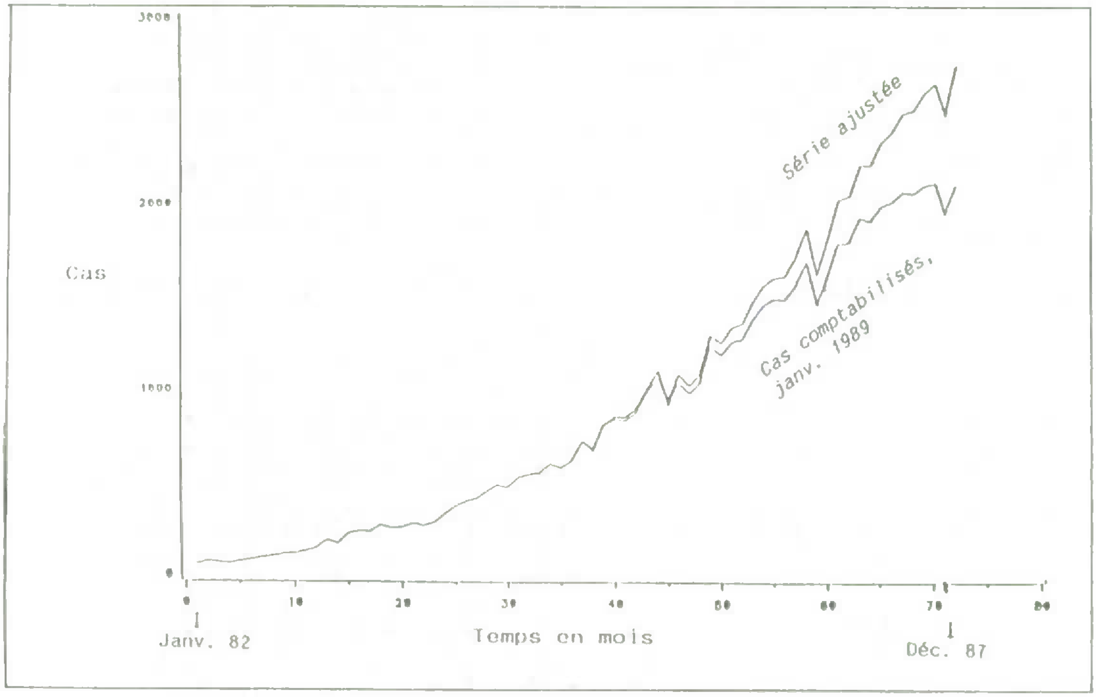


Figure 5. Série ajustée des cas de SIDA diagnostiqués pour cinq dates de déclaration.

où la condition initiale est

$$g(1,s) = \frac{\partial}{\partial t} f_1(s,t)\Big|_{t=0} \quad .$$

MacNeill (1989) présente d'autres équations fonctionnelles ayant cette propriété multiplicative.

données sont considérées inadéquates pour $n = 1, 2, \ldots, 12$. Les estimations ont ensuite été lissées et nous nous sommes servis de (3) pour produire $f(n,\infty)$ dont le graphique est tracé à la figure 3; il a fallu huit termes de (3) pour atteindre la convergence, c'est-à-dire $f(n,60) \equiv f(n,\infty)$. L'application de cet ajustement à la série des cas de SIDA diagnostiqués pour la période de janvier 1982 à décembre 1987, tels que comptabilisés en date de janvier 1989, donne pour résultat la série ajustée représentée graphiquement à la figure 4. L'application de l'ajustement à la série des cas de SIDA diagnostiqués pour chacune des cinq dates de rapport de la figure 1 donne cinq estimations de la série ajustée. Nous pouvons évaluer la validité de l'ajustement ainsi que la stationnarité des conditions de déclaration en nous fondant sur le degre de coincidence des cinq courbes ajustées. La figure 5, où on trouve la représentation graphique des cinq courbes, témoigne de l'efficacité de l'ajustement.

Soulignons également que les équations (1) et (2) sont des applications discontinues d'équations fonctionnelles qui généralisent l'équation fonctionnelle bien connue

$$f(t_1 + t_2) = f(t_1)f(t_2) ,$$

dont la solution, dans des conditions de régularité modérée, est l'exponentielle. La version continue de (2) est l'équation fonctionnelle à deux variables suivante

$$f(s,t_1 + t_2) = f(s,t_1)f(s + t_1,t_2).$$

sous des conditions de régularité modérée, MacNeill (1989), a modélisé les solutions non triviales de la façon suivante:

$$f(s,t) = \exp \{\textstyle\int_0^t g(s + x)dx\}$$

où la condition initiale est

$$g(s) = \frac{\partial}{\partial t}f(s,t)\Big|_{t=0} .$$

L'équation $f(s,t)$ est appelée fonction stationnaire des déclarations tardives. La fonction non-stationnaire des déclarations tardives satisfait l'équation suivante:

$$f_1(s,t_1 + t_2) = f_1(s,t_1)f_{1+t_1}(s + t_1,t_2) .$$

Une fois de plus, sous des conditions de régularité modérée, les solutions non triviales sont modélisées de la façon suivante:

$$f_1(s,t) = \exp \{\textstyle\int_0^t g(1 + x,s + x)dx\}$$

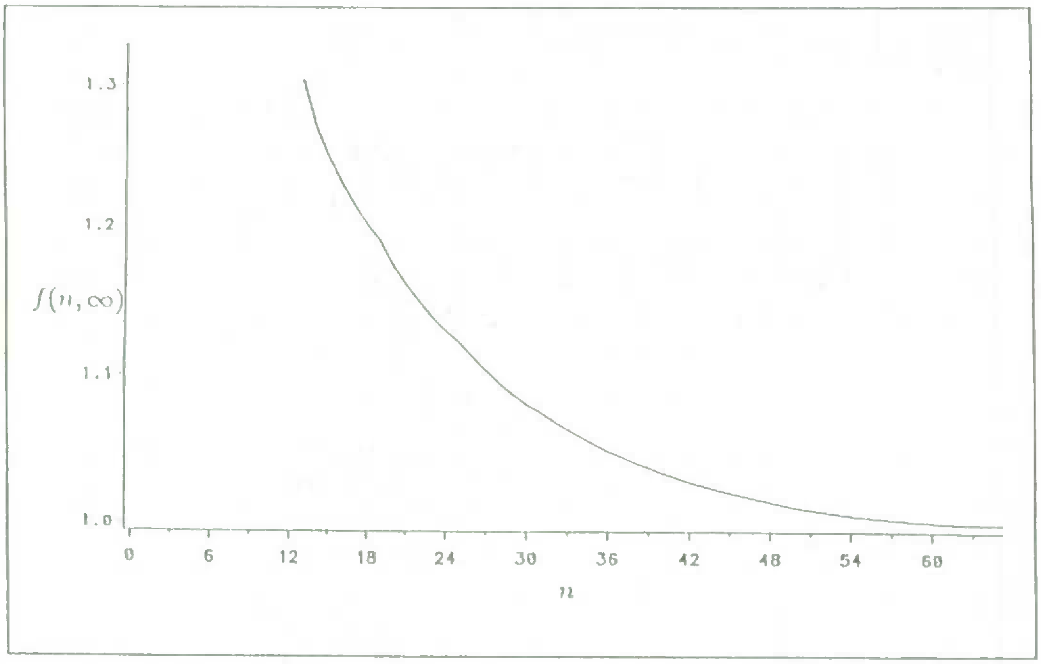


Figure 3. Fonction des déclarations tardives $f(n,\infty)$, $n = 13, 14, \ldots, 60$.

Il est facile de voir que $f_l(n,m)$ satisfait l'équation fonctionnelle suivante:

$$f_l(n,m_1 + m_2) = f_l(n,m_1)f_{l+m_1}(n + m_1,m_2). \tag{1}$$

L'itération de (1) donne

$$f_l(n,m) = \prod_{j=0}^{m-1} f_{l+j}(n + j,1) \quad .$$

Par conséquent, si l'on connaît la condition initiale $f_l(n,1)$ pour tous les n et les l, on obtient $f_l(n,m)$ pour tous les l, n, m. Cela résoudrait le problème des déclarations tardives en situation non stationnaire puisque

$$\hat{D}_{l+\infty}(l - n) = f_l(n,\infty)D_l(l - n).$$

Toutefois, pour pouvoir estimer cette condition initiale, encore faut-il disposer des données nécessaires.

Si $f_{l_1}(n,m) = f_{l_2}(n,m)$ pour toutes les dates de rapport l_1, l_2, alors l'équation fonctionnelle (1) est ramenée à la situation stationnaire qui s'écrit sous la forme:

$$f(n,m_1 + m_2) = f(n,m_1)f(n + m_1,m_2). \tag{2}$$

De même, l'itération de (2) donne pour résultat

$$f(n,m) = \prod_{j=0}^{m-1} f(n + j,1).$$

L'itération de (2) peut aussi donner

$$f(n,6m) = \prod_{j=0}^{m-1} f(n + 6j,6). \tag{3}$$

Cette condition initiale est plus simple et, pourvu que l'efficience de la déclaration soit demeurée relativement constante, elle peut être estimée à partir des données fournies en annexe et illustrées à la figure 1. Pour les données présentées à la figure 1 (et en annexe), l'unité de temps est le mois, où $t = 0$ et $l = 0$, chacun correspondant au mois de décembre 1981. Par exemple, $D_{73}(21) = 264$ est le nombre de cas diagnostiqués en septembre 1983 qui avaient été comptabilisés en date de janvier 1988. Les seules dates de rapport dont nous avons tenu compte sont celles indiquées à la figure 1, c'est-à-dire $l = 61, 67, 73, 79, 85$. Les derniers mois pour lesquels nous disposons de données sont $t = 60$ pour $l = 61$, $t = 66$ pour $l = 67$, etc. A titre d'exemple du coefficient d'ajustement fondé sur les données disponibles, prenons $l = 61$, $m = 12$ et $n = 40$. Alors, $f_{61}(40,\ 12) = 264/259 = 1.0193$.

Le coefficient $f_l(m,n)$ est l'ajustement du nombre de cas de SIDA remontant à n mois dans le passé et comptabilisés au temps l que nous devons appliquer pour reproduire le nombre de cas de SIDA pour le même mois et qui auront été comptabilisés dans les prochains m mois (à partir de l). En supposant des conditions de déclaration d'une efficience constante, alors chacune des fonctions , $f_{61}(n,6)$, $f_{67}(n,6)$, $f_{73}(m.6)$, $f_{79}(n,6)$ servira à estimer le même coefficient. Par exemple, $f_{61}(35,6) = 1.014$, $f_{67}(35,6) = 1.018$, $f_{73}(35,6) = 1.021$ et $f_{79}(35,6) = 1.013$. Morgan et Curran (1986) ont effectué une analyse qui indique que les déclarations tardives aux Etats-Unis n'ont pas changé de manière significative jusqu'en 1986. Donc, en supposant que

$$\hat{f}(n,m) = \underset{l}{\mathrm{ave}}\{f_l(n,m)\}, \tag{4}$$

où l couvre l'ensemble des dates de déclarations pour lesquelles nous pouvons calculer $f_l(n,n)$ à partir des données disponibles. Ainsi, si nous utilisons les données présentées en annexe,

$$\hat{f}(n,6) = \frac{1}{4}\{f_{61}(n,6) + f_{67}(n,6) + f_{73}(n,6) + f_{79}(n,6)\}$$

et, plus précisément, $\hat{f}(35.6) = 1.0161$. Evidemment, $f(n,0) \equiv 1.0$.

Comme nous pouvons le constater dans l'équation (3), seule la condition initiale doit être estimée à partir des données. C'est pourquoi l'équation (4) a uniquement été appliquée à $f(n,6)$ pour $n = 13, 14, \ldots, 60$; les

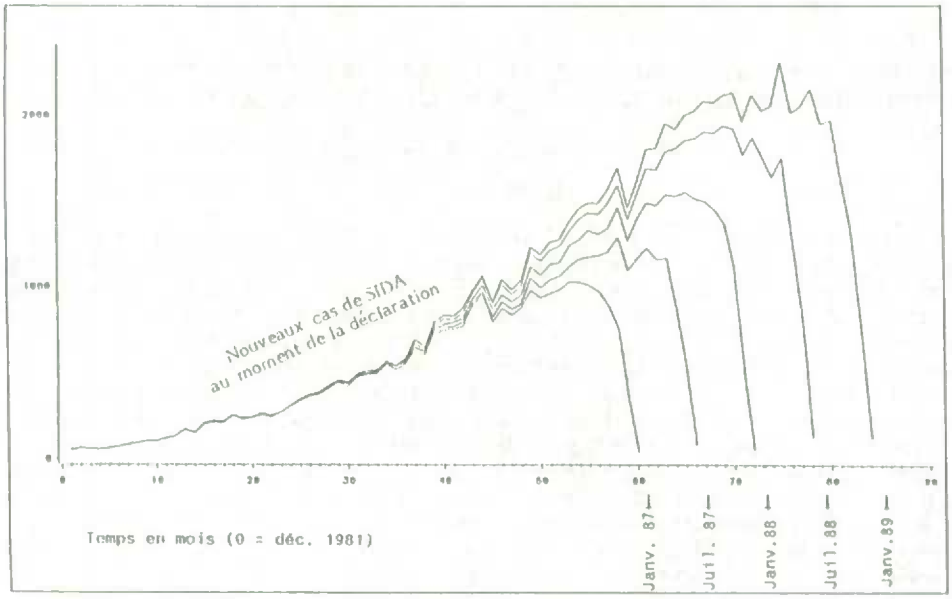


Figure 1. Nombre de cas de SIDA diagnostiqués et signalés au CDC pour certaines dates de déclaration

Le problème dont nous traiterons maintenant est celui qui consiste à estimer le nombre total de cas de SIDA diagnostiqués au cours d'un mois donné malgré les dossiers manquants, comme cela est illustré à la figure 1. Au nombre des méthodes conçues pour tenir compte du problème de déclaration tardive, nous pouvons citer celles de Morgan et Curran (1986) et de Healy et Tillet (1988). Nous proposons une nouvelle méthode.

La figure 2 illustre de façon schématique le problème de déclaration tardive. Supposons que $D_l(t)$ représente le nombre de nouveaux cas de SIDA diagnostiqués au cours de la période de référence t et comptabilisés au temps l. À la figure 2, $D_l(l-n)$ est le nombre de cas comptabilisés maintenant (l) pour un temps remontant à n mois et $D_{l+m}(l-n)$ est le nombre de cas qui auront été comptabilisés dans m mois pour le même mois de diagnostic. L'ajustement pour tenir compte des déclarations tardives est représenté par l'équation

$$f_l(n,m) = \frac{D_{l+m}(l - n)}{D_l(l - n)} .$$

Nous cherchons $f_l(n_l, \infty)$ puisque ce terme représente l'ajustement pour tenir compte des déclarations tardives qui doit être appliqué à $D_l(l - n)$ pour représenter tous les cas diagnostiqués qui seront éventuellement comptabilisés pour le $(l - n)^{ième}$ mois. Cependant, nous ne voulons pas attendre que m devienne aussi élevé.

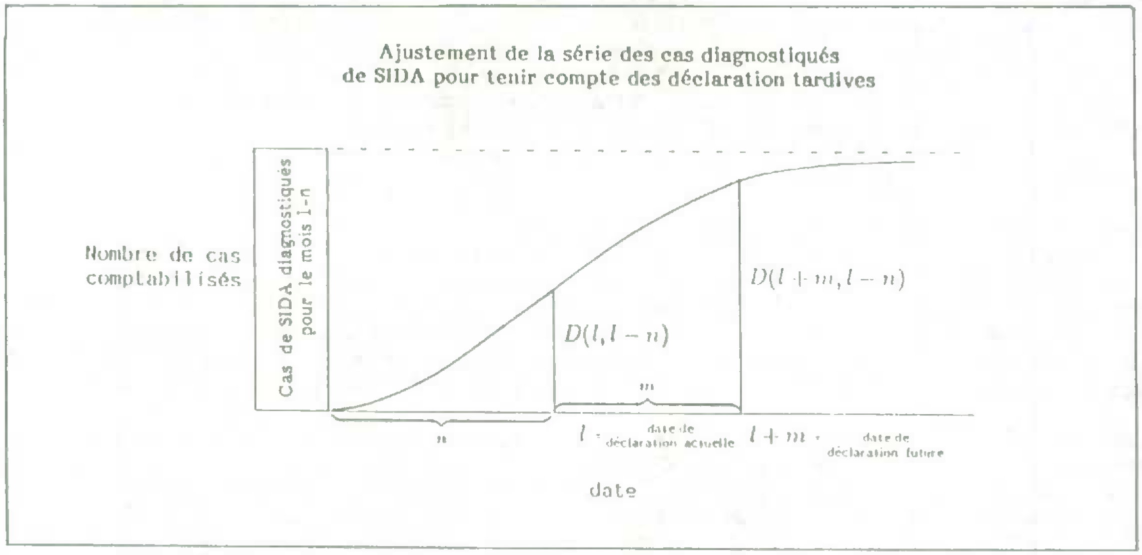

Ajustement de la série des cas diagnostiqués
de SIDA pour tenir compte des déclaration tardives
Nombre de cas comptabilisés
Cas de SIDA diagnostiqués pour le mois l-n
D(l + m, l − n)
D(l, l − n)
m
n
l = date de déclaration actuelle
l + m = date de déclaration future
date

Figure 2. Déclaration tardive des cas diagnostiqués de SIDA

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

AJUSTEMENT POUR TENIR COMPTE DES DÉCLARATIONS TARDIVES DES CAS DE SIDA ET ESTIMATION DE LA POPULATION INFECTÉE PAR LE VIH AUX ETATS-UNIS

I.B. MacNeill[1], Q.P. Duong[2], V.K. Jandhyala[3] et L. Liu[4]

RÉSUMÉ

L'ajustement de la série chronologique des cas diagnostiqués de SIDA pour tenir compte des déclarations tardives est une fonction représentée par un certain nombre d'équations fonctionnelles multiplicatives à plusieurs variables. Nous avons trouvé des solutions à ces équations à la fois pour le cas stationnaire et le cas non stationnaire. L'estimation des conditions initiales est analysée dans le contexte de l'épidémie de SIDA aux Etats-Unis. Nous traitons également du lissage et de l'extrapolation à court terme de la série après ajustement. A la suite d'un examen de la distribution des périodes d'incubation des infections à VIH, nous présentons une équation intégrale qui relie les taux des nouveaux cas diagnostiqués de SIDA aux nouveaux cas d'infection à VIH au moyen de la distribution des périodes d'incubation. Les solutions de cette équation produisent des estimations de la taille de la population infectée par le VIH qui sont plus petites que celles signalées auparavant. La caractéristique la plus importante des estimations de l'infection par le VIH est l'accroissement rapide du taux d'infection avant 1985-1986 et le déclin tout aussi marqué après 1985-1986; le phénomène semble se maintenir malgré des variations importantes dans la distribution des périodes d'incubation et dans les estimations de la taille de la population formée des cas de SIDA diagnostiqués. Nous analysons enfin les répercussions associées à l'évolution à plus long terme de la maladie.

MOTS CLÉS: Ajustement pour tenir compte des déclarations tardives; extrapolation à court terme; équations intégrales; prévision du nombre de cas de SIDA; estimation de l'infection par le VIH.

1. INTRODUCTION

Selon des estimations contenues dans des études réalisées par le bureau du Surgeon General (1986), en 1985, entre 1 et 1.5 million de citoyens américains étaient infectés par le virus de l'immunodéficience humaine (VIH); ces estimations étaient fondées sur de petits échantillons de la population totale. De telles estimations demeurent hautement spéculatives en raison de la difficulté que pose l'obtention de données d'enquête par sondage fiables dans ce domaine. Plus récemment (1989), le bureau du Surgeon General a estimé à 1 million la taille de la population infectée par le VIH, ce qui représente une diminution considérable par rapport aux estimations antérieures, surtout compte tenu de l'intervalle de trois années entre les estimations. Dans la présente étude, nous utilisons des estimations des taux de progression au SIDA des cas d'infection par VIH et des données sur le nombre de cas de SIDA diagnostiqués aux Etats-Unis pour produire de nouvelles estimations du nombre de cas d'infection à VIH. Nous avons mis au point une méthode pour tenir compte des déclarations tardives des cas de SIDA. Les estimations ainsi obtenues ont pour effet de réduire la taille de la population infectée par le VIH, telle qu'évaluée par le bureau du Surgeon General.

2. DÉCLARATIONS TARDIVES

Comme l'épidémie de SIDA se situe maintenant dans une phase cruciale de sa courbe de croissance, il est important, pour les besoins de la prévision à court terme, d'être bien informé du nombre réel de cas de SIDA diagnostiqués chaque mois. Cependant, aux Etats-Unis, une fois posé le diagnostic d'un cas de SIDA, le rapport du cas doit passer par tout un réseau bureaucratique avant de parvenir au Center for Disease Control (CDC). La longueur de l'intervalle de temps qui s'écoule avant que le rapport ne soit reçu au CDC varie. Dans un petit nombre de cas seulement, le rapport sera produit au cours d'un mois donné sera dans les mains du CDC dans les quelques mois suivants. La plupart des cas de SIDA sont comptabilisés dans les douze mois qui suivent le diagnostic. Cependant, certains cas peuvent être comptabilisés comme ayant été diagnostiqués au cours d'un mois donné après un intervalle aussi long que plusieurs années; un problème que vient accentuer les changements de définition de la maladie.

Ce problème de déclaration tardive est illustré par les données du graphique de la figure 1. Ces données sont tirées de rapports semestriels publiés par le CDC, soit ceux datés de janvier 1987, de juillet 1987, de janvier 1988, de juillet 1988 et de janvier 1989, et elles représentent les chiffres mensuels des cas de SIDA diagnostiqués et comptabilisés au CDC au moment de la production du rapport. Les chiffres récents se rapportant aux cas de SIDA diagnostiqués il y a longtemps semblent représenter la quasi-totalité des cas diagnostiqués puisque les cinq courbes sont très rapprochées les unes des autres, mais les chiffres relatifs aux cas diagnostiqués plus récemment font ressortir de façon évidente la longueur des délais de déclaration.

[1] Department of Statistical and Actuarial Sciences and Epidemilogy and Biostatistics, The University of Western Ontario, London, (Ontario), Canada N6A 5B9

[2] Bureau of Management Consulting, 364 Laurier Avenue W., Ottawa, (Ontario), Canada K1A 0S5

[3] Department of Mathematics, Washington State University, Pullman Washington, U.S.A. 99163

[4] Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, (Ontario) Canada N6A 5B9

SECTION 5

ÉPIDÉMIOLOGIE

BIBLIOGRAPHIE

Alba, E. de (1988), "Temporal Disaggregation of Time Series: a Bayesian Analysis", **Journal of Business and Economic Statistics**, Vol. 6, No. 2, pp 197-206.

Bournay, J., Laroque, G. (1979), «Réflexions sur la méthode d'élaboration des comptes trimestriels», **Annales de l'I.N.S.É.É.**, Vol. 36, pp. 3-30.

Boot, J.C.G., Feibes, W., and Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data", **Applied Statistics**, Vol. 16, no. 1, pp. 65-75

Cholette, P.A. (1988), "Weights to Calendarize Fiscal year Data Referring to Any Consecutive 12 Months or 4 Quarters", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-88-023E.

Cholette, P.A., Baldwin, A. (1989), "Converting Fiscal Year Data into Calendar Year Values", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-89-007E; soumis pour publication dans **The Journal of Business of Economic Statistics**.

Cholette, P.A., Chhab N. (1988), "Converting Aggregates of Weekly Data into Monthly values", **Statistics Canada**, Time Series Research and Analysis Division, Research Paper No. TSRA-89-019E; soumis pour publication dans **Applied Statistics**.

Cholette, P.A, Dagum, E. Bee (1989), "Benchmarking Socio-Economic Time Series Data: A Unified Approach", **Statistics Canada**, Time Series Research and Analysis Division, Working Paper No. 89-006E.

Chow, G.C., Lin, An-Loh (1971), "Best linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series", **Review of Economics and Statistics**, Vol. 53, No. 4, pp. 372-375.

Cohen, K.J., Müller, W., and Padberg, M.W. (1971) "Autoregressive Approaches to the Disaggregation of Time Series Data", **Applied Statistics**, Vol. 20, pp 119-129.

Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", **Journal of the American Statistical Association**, Vol. 66, No. 333, pp. 99-102.

Dagum, E. (1980), The X-11-ARIMA Seasonal Adjustment Programme, Statistics Canada, Cat. 12-564E; La Méthode de désaisonnalisation X-11-ARIMA, **Statistique Canada**, Cat. 12-564F.

Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series", **Review of Economic and Statistics**, Vol. 63, pp. 471-476.

Helfand, S.D., Monsour, J.J., Trager, M.L., "Historical Revision of Current Business Survey Estimates", **Proceedings of the Business and Economic Statistics Section**, American Statistical Association, 1977.

Smith, P. (1977), "Alternative Methods for Step-Adjustment", **Statistics Canada**, Econometrics Section, Current Economic Analysis Division (internal document).

Young, A.H. (1965), "Estimating Trading-Day Variations in Monthly Economic Time Series", U.S. Bureau of the Census, Technical Paper No. 12.

Tableau 3

Analyse et comparaison des erreurs pourcentuelles absolues de trimestrialisation

Tableau 3A: pour les 11 premiers trimestres civils des séries, en utilisant la méthode proposée et en assimilant les trimestres financiers au trimestre civil le plus proche entre parenthèses

	Moyenne		écarts types		minimums		maximums		σ_s	σ_I
Épiceries et boucheries	0.6	(1.3)	0.4	(0.6)	0.0	(0.4)	1.1	(2.3)	5.9	1.1
Grand magasins	0.3	(6.2)	0.2	(3.0)	0.0	(1.6)	0.8	(10.3)	27.6	1.6
Magasins de marchandises diverses	1.1	(7.2)	0.7	(3.4)	0.2	(2.2)	2.6	(12.5)	18.0	3.1
Magasins généraux	0.5	(3.2)	0.3	(1.7)	0.0	(0.4)	1.2	(6.5)	10.8	1.8
Bazars	1.5	(5.4)	1.6	(4.6)	0.1	(0.3)	4.6	(12.1)	29.1	2.3
Concessionnaires d'automobiles	0.7	(7.6)	0.3	(6.2)	0.2	(0.1)	1.2	(17.6)	14.1	3.6
Vendeur d'auto. d'occasion	0.8	(6.9)	0.8	(6.5)	0.0	(0.4)	2.9	(17.5)	15.8	4.5
Stations-services	0.4	(2.5)	0.2	(1.5)	0.0	(0.5)	0.7	(5.3)	7.0	1.4
Garages	0.6	(3.3)	0.5	(1.9)	0.0	(0.5)	1.5	(6.9)	7.5	2.6
Magasins de pièces d'automobiles	1.0	(6.8)	0.8	(5.2)	0.0	(0.3)	2.3	(15.1)	20.5	2.8

Tableau 3B: EPA pour le dernier trimestre civil de chaque en utilisant la méthode proposée

Épiceries et boucheries	0.6
Grands magasins	0.4
Magasins de marchandises diverses	2.4
Magasins généraux	0.6
Bazars	1.1
Concessionnaires d'automobiles	1.9
Vendeurs d'automobiles d'occasion	0.7
Stations-services	1.4
Garages	1.3
Magasins de pièces d'automobiles	0.2

7. RELATION AVEC LES MÉTHODES APPARENTÉES

Comme expliqué dans la section 3, la méthode de trimestrialisation exposée ici est une adaptation des méthode d'étalonnage du type Denton (ex.: Denton, 1971, Helfand, Monsour et Trager, 1977). L'adaptation consiste simplement à permettre aux jalons de couvrir des trimestres financiers, au lieu d'année civiles. Il s'agit ensuite d'ajuster (d'étalonner) un profil saisonnier et de rotation des jours aux jalons trimestriels financiers. Les valeurs trimestrialisées sont ensuite posées égales aux sommes sur trimestres civils de la série étalonnée. Cholette et Baldwin (1989) ont proposé la même stratégie pour annualiser des chiffres d'années financières; et Cholette et Chhab (1989), pour mensualiser des agrégats de chiffres hebdomadaires. La variante logarithmique de la section 3 peut se voir comme une approximation de la variante proportionnelle de Denton (1971), souvent utilisée d'ailleurs comme une approximation à une variante de taux de croissance (Smith, 1977).

Il existe une littérature au sujet de la «désagrégation temporelle». La méthode de Boot, Feibes et Lisman (1967) coïncide avec la variante de la section 3, si les jalons couvrent des années civiles et si le profil saisonnier est trimestriel et égal à zéro. Cette méthode s'utilise pour transformer des chiffres annuels civils en valeurs trimestrielles sans saisonnalité. Cohen, Müller et Padberg (1971) ont généralisé l'approche pour transformer des données *civiles* de n'importe quelle fréquence en valeurs plus fréquentes sans saisonnalité.

Les méthodes de désagrégation temporelles proposées par Chow et Lin (1971), Bournay et Laroque (1979), Fernandez (1981), de Alba (1988) et d'autres interpolent entre jalons en utilisant des séries apparentées dans le cadre d'une régression linéaire. Ces méthodes coïncident avec la variante additive ci-dessus, si les jalons reflètent des périodes civiles, si on n'a qu'un seul régresseur à coefficient unité et si le coefficient d'auto-corrélation des résidus de la régression est posé égal à 1.

8. CONCLUSION

Ce travail a proposé une méthode pour transformer les chiffres de trimestres financier en valeurs de trimestres civils. L'application de la méthode a quelques séries a produit des résultats encourageants. Il serait désirable de poursuivre des recherches, notamment en ce qui concerne les estimations préliminaires et l'utilisation de profils saisonniers approximatifs.

Comme expliqué, l'omission de la trimestrialisation conduit à des chiffres erronés. Pourtant, à notre connaissance, le problème n'a pas jusqu'à ce jour retenu l'attention des statisticiens.

plupart des grands magasins ont des années financières se terminant en janvier et que le régime de trimestres financiers de la figure correspond à cette année financière.

Les tableaux 2 et 3 présentent les résultats pour les dix séries examinées. Le tableau 2A consigne des statistiques sur les erreurs pourcentuelles absolues (EPA) des 36 interpolations, par rapport aux vraies valeurs mensuelles. Les faibles valeurs des moyennes et des écarts-types des EPA révèlent, dans plusieurs cas, un degré surprenant de précision. Ceci démontre la possibilité d'obtenir des interpolations infra-annuelles assez précises à partir d'un simple profil saisonnier et de jalons. Le tableau consigne également les écarts-types du profil saisonnier et de la composante irrégulière, σ_s et σ_I, estimés par X-11-ARIMA. La précision (faibles statistiques) est négativement corrélée à l'intensité de la composante irrégulière (mesurée par σ_I) de chaque série, ce qui n'a rien d'étonnant.

Le tableau 2B présente les mêmes statistiques pour les trois dernières interpolations. Contrairement aux attentes (voir section 5.3), celles-ci n'apparaissent pas ostensiblement moins précises que les autres. Ceci est dû au fait que, pendant le quatrième trimestre de 1988, aucune des séries n'affiche de revirement de tendance-cycle. Cependant; les EPA minimums diffèrent entre les tableaux 2A et 2B, indiquant que pour chacune des séries la précision ne culmine jamais durant les trois derniers mois. Au contraire, les EPA maximums coïncident pour 3 des 10 séries, indiquant que la précision atteint souvent son minimum durant les trois derniers mois.

Le tableau 3A présente les statistiques relatives aux EPA des valeurs trimestrialisées, obtenues (1) par la méthode proposée et (2) en assimilant, sans correction, les données financières aux trimestre qui chevauche le plus, entre parenthèses. Les moyennes des EPA sont de deux à vingt fois plus basses avec la méthode proposée. La réduction est particulièrement frappante pour certaines séries à forte saisonnalité (mesurée par σ_s), ce qui n'est pas surprenant.

Le tableau 3B consigne les EPA pour l'estimation du dernier trimestre civil des dix séries, selon la méthode proposée. La discussion ci-dessus entourant le tableau 2B reste pertinente.

Les résultats obtenus ici sont plutôt encourageants pour la méthode proposée. Cependant, en pratique, le profil saisonnier ne serait pas connu avec autant de précision. Les résultats obtenus ici pourraient donc s'interpréter comme un échantillon des meilleurs résultats qu'on puisse espérer dans les cas véritables de trimestrialisation.

Tableau 2
Analyse des erreurs pourcentuelles absolues d'interpolations

Tableau 2A: pour les 36 observations des séries

	moyenne	ec. type	min.	max.	σ_s	σ_I
Épiceries et boucheries	0.8	0.7	0.1	2.6	5.9	1.1
Grand magasins	0.8	0.7	0.1	2.3	27.6	1.6
Magasins de Marchandises diverses	2.6	1.9	0.1	7.2	18.0	3.1
Magasins généraux	1.0	0.7	0.0	3.0	10.8	1.8
Bazars	2.1	2.7	0.0	12.5	29.1	2.3
Concessionnaires automobiles	2.0	1.5	0.2	6.8	14.1	3.6
Vendeurs d'auto usagées	2.0	2.2	0.0	10.1	15.8	4.5
Stations-services	0.7	0.6	0.0	2.6	7.0	1.4
Garages	1.1	0.8	0.0	3.2	7.5	2.6
Magasins de pièces d'automobiles	1.9	1.5	0.1	5.7	20.5	2.8

Tableau 2B: pour les trois derniers mois des séries

	moyenne	ec. type	min.	max.
Épicerie et boucheries	1.1	0.3	0.8	1.5
Grand magasins	0.5	0.3	0.3	1.0
Magasins marchandises diverses	2.4	0.9	1.2	3.5
Magasins généraux	1.6	1.1	0.2	3.0
Bazars	2.1	1.6	0.9	4.4
Concessionnaires d'automobiles	3.1	2.6	1.2	6.8
Vendeurs d'auto usagées	1.9	0.9	0.6	3.0
Stations-services	1.4	0.9	0.7	2.6
Garages	1.4	1.1	0.3	2.8
Magasins de pièces d'automobiles	3.5	1.5	1.9	5.5

5.4 Application lorsque M < 4

Lorsqu'un seul trimestre financier est disponible, couvrant de février à avril 86 disons, on utilise les poids du tableau 1A (a). Ces poids distribuent simplement 1/3 de l'écart R_1 sur le profil saisonnier S. Les interpolations pour les mois de janvier à juin 86 sont alors parfaitement parallèles à S et contiennent seulement de la saisonnalité (à moins que S ne soit pas seulement saisonnier). On peut démontrer que les valeurs trimestrialisées C_n (n=1,2) pour les premier et deuxième trimestres de 86 sont triviales et égales à F_1. La trimestrialisation a donc peu de chance de réussir en présence d'une seule valeur financière.

Lorsque le second trimestre financier de 86 devient disponible, on applique les poids du tableau 1B (a) aux données de l'intervalle janvier à septembre 86. Les interpolations sont supérieures à ce qu'elles étaient, et des valeurs trimestrialisées non triviales C_n (n=1,2,3) sont maintenant obtenues pour les trois premiers trimestres de 86. Les estimations triviales des deux premiers trimestres obtenues antérieurement sont révisées. De manière analogue, lorsque le troisième trimestre financier de 86 devient disponible, on applique les poids du tableau 1C aux données de l'intervalle janvier à décembre 86. Toutes les estimations obtenues antérieurement sont révisées.

6. EXEMPLES DE LA VARIANTE LOGARITHMIQUE

Afin de mettre à l'épreuve l'approche proposée, la présente section applique la variante logarithmique de la section 4 à dix séries mensuelles du Commerce de détail canadien. La méthode est appliquée de manière mobile sur 5 trimestres civils comme expliqué à la section 5. Les séries, s'étendant de janvier 1986 à décembre 88, furent converties en chiffres de trimestres financiers, couvrant février à avril, mai à juillet, etc. Les chiffres mensuels furent ensuite récupérées, en tant qu'interpolations, en appliquant la méthode aux jalons financiers, F, et à un profil saisonnier et de rotation des jours, S. Pour chaque série, S avait été calculé par la méthode de désaisonnalisation X-11-ARIMA (Dagum, 1980), appliquée aux chiffres mensuels. Normalement, S proviendrait d'une autre source.

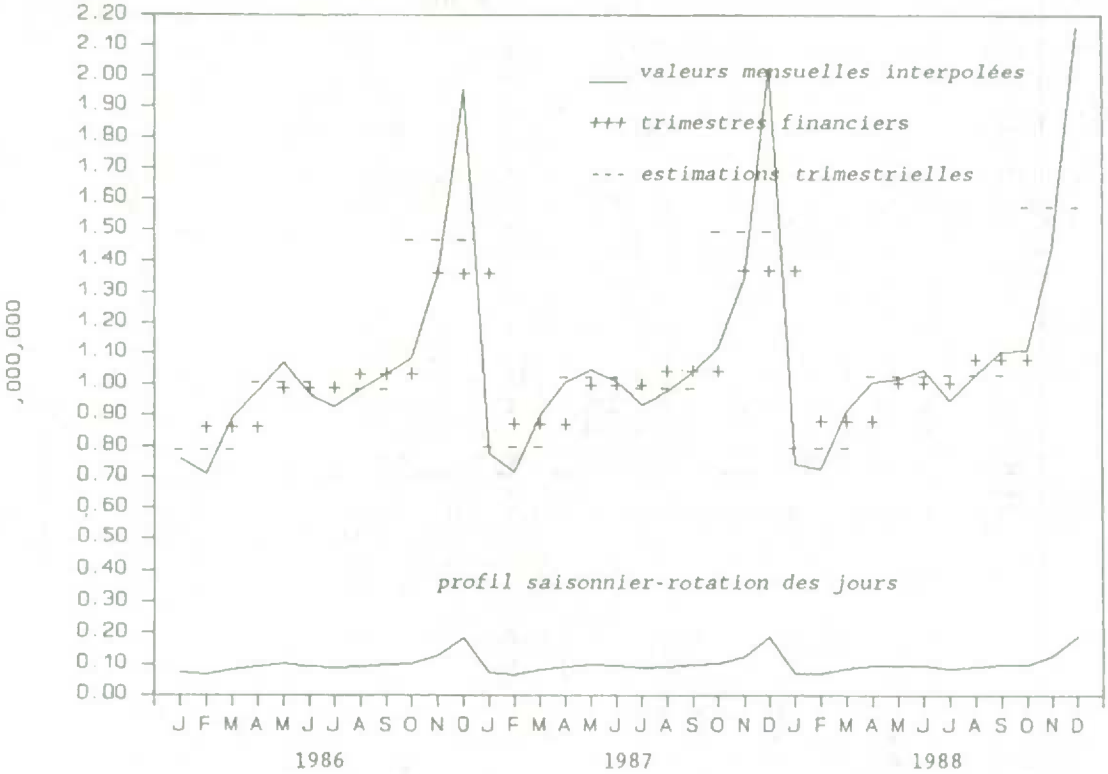


Figure 2: Estimations mensuelles interpolées et estimations trimestrielles civiles (---) obtenues en appliquant la variante logarithmique aux données de trimestres financiers (+++) et au profil saisonnier illustré

La figure 2 illustre le cas des Grand magasins. Les interpolations Γ^*_t adoptent les taux de croissance d'un mois à l'autre du profil saisonnier et se conforment exactement aux jalons trimestriels financiers F_m. Cette conformité se vérifie également pour les estimations trimestrielles civiles, étant les sommes trimestrielles des interpolations. À remarquer que la

Tableau 1

Poids $W_{t,m}$ appliqués aux écarts trimestriels R_m pour obtenir les interpolations mensuelles, selon les régimes réguliers de trimestres financiers des colonnes (a), (b) et (c)

Tableau 1A: lorsqu'on a un seul trimestre financier (M=1)

régime trimestriel:	(a)	(b)	(c)
t\m	F M A	M A M	J F M
J	0.33333	0.33333	0.33333
F	0.33333	0.33333	0.33333
M	0.33333	0.33333	0.33333
A	0.33333	0.33333	0.33333
M	0.33333	0.33333	0.33333
J	0.33333	0.33333	0.33333

Tableau 1B: lorsqu'on deux trimestres financiers (M=2)

régime trimestriel:	(a)		(b)		(c)	
t\m	F M A	M J J	M A M	J J A	J F M	A M J
J	0.40351	-0.07018	0.40351	-0.07018	0.40351	-0.07018
F	0.40351	-0.07018	0.40351	-0.07018	0.35088	-0.01754
M	0.35088	-0.01754	0.40351	-0.07018	0.24561	0.08772
A	0.24561	0.08772	0.35088	-0.01754	0.08772	0.24561
M	0.08772	0.24561	0.24561	0.08772	-0.01754	0.35088
J	-0.01754	0.35088	0.08772	0.24561	-0.07018	0.40351
J	-0.07018	0.40351	-0.01754	0.35088	-0.07018	0.40351
A	-0.07018	0.40351	-0.07018	0.40351	-0.07018	0.40351
S	-0.07018	0.40351	-0.07018	0.40351	-0.07018	0.40351

Tableau 1C: lorsqu'on a trois trimestres financiers (M=3)

	régime trimestriel (a)			régime trimestriel (b)		
t\m	F M A	M J J	A S O	M A M	J J A	S O N
J	0.40676	-0.08889	0.01546	0.40676	-0.08889	0.01546
F	0.40676	-0.08889	0.01546	0.40676	-0.08889	0.01546
M	0.35169	-0.02222	0.00386	0.40676	-0.08889	0.01546
A	0.24155	0.11111	-0.01932	0.35169	-0.02222	0.00386
M	0.07633	0.31111	-0.05411	0.24155	0.11111	-0.01932
J	-0.02222	0.37778	-0.02222	0.07633	0.31111	-0.05411
J	-0.05411	0.31111	0.07633	-0.02222	0.37778	-0.02222
A	-0.01932	0.11111	0.24155	-0.05411	0.31111	0.07633
S	0.00386	-0.02222	0.35169	-0.01932	0.11111	0.24155
O	0.01546	-0.08889	0.40676	0.00386	-0.02222	0.35169
N	0.01546	-0.08889	0.40676	0.01546	-0.08889	0.40676
D	0.01546	-0.08889	0.40676	0.01546	-0.08889	0.40676

Tableau 1D: lorsqu'on a quatre trimestres financiers ou plus ($M \geq 4$)

	régime trimestriel (a)				régime trimestriel (b)			
t\m	F M A	M J J	A S O	N D J	M A M	J J A	S O N	D J F
J	0.40692	-0.08980	0.01962	-0.00341	0.40692	-0.08980	0.01962	-0.00341
F	0.40692	-0.08980	0.01962	-0.00341	0.40692	-0.08980	0.01962	-0.00341
M	0.35173	-0.02245	0.00491	-0.00085	0.40692	-0.08980	0.01962	-0.00341
A	0.24135	0.11225	-0.02453	0.00427	0.35173	-0.02245	0.00491	-0.00085
M	0.07577	0.31430	-0.06868	0.01194	0.24135	0.11225	-0.02453	0.00427
J	-0.02245	0.37909	-0.02821	0.00491	0.07577	0.31430	-0.06868	0.01194
J	-0.05332	0.30662	0.09689	-0.01685	-0.02245	0.37909	-0.02821	0.00491
A	-0.01685	0.09689	0.30662	-0.05332	-0.05332	0.30662	0.09689	-0.01685
S	0.00491	-0.02821	0.37909	-0.02245	-0.01685	0.09689	0.30662	-0.05332
O	0.01194	-0.06868	0.31430	0.07577	0.00491	-0.02821	0.37909	-0.02245
N	0.00427	-0.02453	0.11225	0.24135	0.01194	-0.06868	0.31430	0.07577
D	-0.00085	0.00491	-0.02245	0.35173	0.00427	-0.02453	0.11225	0.24135
J	-0.00341	0.01962	-0.08980	0.40692	-0.00085	0.00491	-0.02245	0.35173
F	-0.00341	0.01962	-0.08980	0.40692	-0.00341	0.01962	-0.08980	0.40692
M	-0.00341	0.01962	-0.08980	0.40692	-0.00341	0.01962	-0.08980	0.40692

5.2 Application lorsque $M \geq 4$

Lorsque c'est possible, nous recommandons l'application des poids à la manière d'une moyenne mobile de 4 trimestres financiers ($M=4$) encastrés dans 5 trimestres civils. Pour fins d'illustration, on suppose une série débutant en janvier 1986, comprenant initialement 4 trimestres financiers, couvrant février à avril, mai à juillet, etc., encastrés dans 5 trimestres civils. À mesure que des données deviennent disponibles, la série comprend éventuellement 11 trimestres financiers encastrés dans 12 trimestres civils. Les poids du tableau 1D (a) s'appliquent alors 8 fois de la manière suivante.

La première fois, les poids s'appliquent aux 4 premiers chiffres financiers F_m ($m=1,\ldots,4$) et aux 15 chiffres saisonniers S_t ($t=1,\ldots 15$) de l'intervalle janvier 86 à mars 87. Par exemple les interpolations additives de mars et avril 86 sont

$$\Gamma^*_3 = S_3 + 0.35173\ R_1 - 0.02245\ R_2 + 0.00491\ R_3 - 0.00085\ R_4,$$

$$\Gamma^*_4 = S_4 + 0.24135\ R_1 + 0.11225\ R_2 - 0.02453\ R_3 + 0.00427\ R_4.$$

où les R_m sont les écarts de (5.1) entre F_m et S_t. Ceci produit les estimations finales pour janvier à septembre 86 (et pour les trimestres civils correspondants) et les estimations préliminaires pour janvier à mars 87.

La deuxième fois, c-à-d lorsque le 5^e chiffre trimestriel financier devient disponible, les mêmes poids s'appliquent aux données F et S de l'intervalle avril 86 à juin 87. Ceci produit les estimations finales pour octobre à décembre 86 (et pour les`trimestres civils correspondants); les estimations révisées pour janvier à mars 87 et les estimations préliminaires pour avril à juin 87.

La troisième fois, les poids s'appliquent aux données F et S de l'intervalle juillet 86 à septembre 87. Ceci produit les estimations finales pour janvier à mars 87, les estimations révisées pour avril à juin 87 et les estimations préliminaires pour juillet à septembre 87.

Et ainsi de suite. Cette application des poids, sur intervalles mobiles de cinq trimestres civils, réduit à deux le nombre de révisions; et assure que les estimations civiles finales aient deux jalons financiers «de chaque côté» et soient ainsi centrales dans chaque intervalle. Cette mise en oeuvre est aussi beaucoup plus économique que d'appliquer la méthode (recalculer W) sur toutes les données disponibles et n'affecte pas les estimations de façon notable.

5.3 Fiabilité des estimations préliminaires

Les estimations préliminaires sont sujettes à des révisions plus importantes que les estimations révisées (une fois), parce que certains des mois impliqués sont à l'extérieur de l'intervalle couvert par les jalons financiers. Sous le régime financier de la colonne (a), les deux derniers mois sont à l'extérieur; sous le régime (b), le dernier mois; et sous le régime (c), les trois derniers mois. Nous recommandons par conséquent la non-utilisation des estimations préliminaires, spécialement si on anticipe un revirement (à la baisse ou à la hausse) dans le cycle des affaires.

Les statisticiens incapables de tolérer le délai de production résultant (ou la fiabilité réduite) pourraient élaborer une prévision du prochain jalon financier et appliquer la méthode à la série artificiellement prolongée. (Ceci tend à améliorer les résultats pour certaines méthodes de désaisonnalisation, Dagum, 1980.) Un bon point de départ pour cette prévision est $F^f_m = F_{m-1} + F_{m-4} - F_{m-5}$. Cette prévision est celle d'un modèle ARIMA dégénéré (0,1,0)(0,1,0). Ce modèle pose que le changement d'un trimestre à l'autre tend à se répéter d'une année à l'autre, ce qui implique une saisonnalité constante et une tendance-cycle localement *linéaire* durant les quatre derniers trimestres.

Pour les séries de flux, la solution (4.1) préserve aussi les taux de croissance, mais les interpolations ont leurs produits trimestriels financiers égaux aux jalons financiers. Pour obtenir les égalités de sommes, une stratégie fructueuse consiste à itérer sur les valeurs de ln F. D'excellentes valeurs de départ pour ln $F^{(1)}$ proviennent des produits trimestriels financiers de S multipliés par les écarts proportionnels entre F et S (entre crochets):

$$F_m^{(1)} = \prod_{t\in m} S_t \, [F_m /(\sum_{\tau\in m} S_\tau)] \quad \Rightarrow \quad \ln F_m^{(1)} = \sum_{t\in m} \ln S_t \, [F_m /(\sum_{\tau\in m} S_\tau)], \quad m=1,\dots,M. \tag{4.2}$$

Les premières interpolation $\Gamma^{*(1)}$ s'obtiennent au moyen de (4.1) appliqué à ln $F^{(1)}$ de (4.2). Pour les autres itérations (k>1), les valeurs révisées de ln $F^{(k)}$ proviennent du produit de $F^{(k-1)}$ et des écarts proportionnels résiduels entre F and $\Gamma^{*(k-1)}$ (entre crochets):

$$F_m^{(k)} = F_m^{(k-1)} \, [F_m /(\sum_{\tau\in m} \Gamma^{*}_{\tau}{}^{(k-1)})] \Rightarrow \ln F_m^{(k)} = \ln F_m^{(k-1)} + \ln [F_m /(\sum_{\tau\in m} \Gamma^{*}_{\tau}{}^{(k-1)})], \quad m=1,\dots M. \tag{4.3}$$

Les interpolations ultérieures $\Gamma^{*(k)}$ (k>1) s'obtiennent de (4.1) appliquée à ln $F^{(k)}$ de (4.3). Les itérations entre (4.3) et (4.1) se poursuivent jusqu'à satisfaction des égalités (3.1b) (où $\epsilon=0$) à plus de 0.25% (disons), ce qui nécessite généralement moins de 5 itérations (K<5). Une réalisation exacte de (3.1b) peut s'obtenir en multipliant les dernières interpolations $\Gamma^{*(K)}$ par les derniers écarts proportionnels résiduels:

$$\Gamma^{*}_{t} = \Gamma^{*}_{t}{}^{(K)} \, [F_m /(\sum_{t\in m} \Gamma^{*}_{t}{}^{(K)})] \tag{4.4}$$

Les avantages de la variante logarithmique sont les suivants:
(1) S peut avoir un ordre de grandeur différent de F; et
(2) les poids W de (3.8) sont calculés une fois pour toute et peuvent s'appliquer à n'importe quelles données S et F, quelles qu'en soient les valeurs.
Ces propriétés combinent les avantages des variantes proportionnelle et additive de l'étalonnage de type Denton.

5. MISE EN OEUVRE

Dans la variante additive, comme dans la variante logarithmique, les interpolations mensuelles sont égales au profil saisonnier (et de rotation des jours) choisi, plus une combinaison linéaire des écarts R=F-BS entre les jalons trimestriels financiers et les sommes correspondantes du profil saisonnier:

$$\Gamma^{*} = S + W [F - B S] = S + W R \quad \Rightarrow \quad \Gamma^{*}_{t} = S_t + \sum_{m=1}^{M} W_{t,m} R_m. \tag{5.1}$$

$$\begin{aligned} \ln \Gamma^{*(k)} &= \ln S + W [\ln F^{(k)} - B \ln S] = \ln S + W R^{(k)} \\ \Rightarrow \ln \Gamma^{*}_{t}{}^{(k)} &= \ln S_t + \sum_{m=1}^{M} W_{t,m} R^{(k)}_{m}. \end{aligned} \tag{5.2}$$

où (5.1) s'applique pour la variante additive et (5.2) pour la variante logarithmique, et ou les poids $W_{t,m}$ sont donnés par (3.8) (avec $V_\epsilon=0$).

5.1 Description des poids

Le tableau 1 contient les poids $W_{t,m}$ pour les trois régimes réguliers de trimestres financiers possibles, dans lesquels chaque trimestre contient trois mois. (Il arrive parfois qu'un trimestre financier compte plus ou moins que 3 mois.) Le tableau 1A contient les poids à utiliser, lorsqu'on a qu'un trimestre financier, M=1; le tableau 1B, lorsqu'on a deux trimestres financiers, M=2; le tableau 1C, M=3; et le tableau 1D, M=4.

La colonne (a) de chaque sous-tableau se rapporte au régime financier où les trimestres couvrent de février à avril, mai à juillet, etc. (ou bien de mai à juillet, août à octobre, etc.); la colonne (b), au régime où les trimestres couvrent de mars à mai, juin à août, etc.; et la colonne (c), au régime où les trimestres sont civils. Les poids de la colonne (c) peuvent s'utiliser pour interpoler des valeurs mensuelles à partir de trimestres civils. À cause du manque d'espace, la colonne (c) n'apparait pas dans les sous-tableaux C et D. Cependant, on peut facilement la reconstruire à partir de la colonne (b): la rangée t de la colonne (c) répète la rangée t-1 de la colonne (b); en fait la même relation s'observe entre les rangées des colonnes (b) et (a). En conséquence de cette relation, il suffit d'emmagasiner les poids des colonnes (a) seulement.

La solution (3.8) nécessite une inversion de matrice beaucoup plus petite que (3.5). Cependant, la solution (3.8) ne serait pas pertinente, si la variance (3.6) est calculée, ce qui nécessite la grande inversion de (3.5).

Contrairement à (3.5), la solution (3.8) admet $V_\epsilon=0$. À moins d'indication contraire, le reste du document suppose $V_\epsilon=0$. Lorsque $V_\epsilon=0$, la valeur de σ_u^2 implicite dans V_* n'a aucune importance, car elle s'annule; de même $1-\rho^2$ s'annule. En outre, avec $V_\epsilon=0$, (3.8) a la même forme que la solution de Denton (1971); et (3.8) est également la solution obtenue en minimisant la fonction objective contrainte suivante:

$$\sum_{t=2}^{T} ((\Gamma_t-S_t) - \rho(\Gamma_{t-1}-S_{t-1}))^2 \quad - 2 \sum_{m=1}^{M} \lambda_m [(\sum_{\tau\in m} \Gamma_\tau) - F_m], \qquad (\Gamma_t-S_t=e_t).$$

Comme signalé par Bournay et Laroque (1979) pour l'étalonnage, à mesure que ρ tend vers 1, cette fonction tend vers celle minimisée par Denton et d'autres (sauf pour les contraintes); elle spécifie que Γ préserve le changement d'un mois à l'autre observé pour S.

3.3 Les valeurs trimestrialisées

Que les interpolation proviennent de (3.8) ou de (3.5), les estimations de trimestres civils désirées sont simplement données par les sommes appropriées de Γ^*:

$$C^* = G\ \Gamma^*, \qquad G = I_N \otimes [1\ 1\ 1]. \tag{3.9}$$

où N est le nombre de trimestres civils et ou I_N est la matrice identité N par N. La variance de C^* s'obtient de celle de Γ^*:

$$\mathrm{var}(C^*) = G\ \mathrm{var}(\Gamma^*)\ G'. \tag{3.10}$$

Si on ne s'intéresse pas aux interpolations mensuelles en tant que telles, les estimations peuvent s'exprimer directement en fonction des données de base F et S, par substitution de (3.8) dans (3.9):

$$C^* = G\ (S + W\ [\ F - B\ S\ ]) \ = \ G\ S + P\ [\ F - B\ S\ ]. \tag{3.11}$$

Les poids W de (3.8) et P de (3.11) ne dépendent aucunement des données F et S. Ils dépendent seulement de la longueur T de la série et du régime de trimestres financiers, à savoir si les trimestres se terminent en janvier, avril, juillet, etc., ou bien en février, mai, août, etc. On peut donc considérer les poids comme connus d'avance et les appliquer à toute série ayant même longueur et même régime financier. Comme expliqué à la section 5, ceci aura des avantages importants au niveau de la mise en oeuvre de la méthode.

4. LA VARIANTE LOGARITHMIQUE

La méthode additive présentée dans la section 3 convient lorsque le profil saisonnier-rotation des jours S est du même ordre de grandeur que les chiffres des trimestres financiers (divisés par 3). Or S s'exprime plus aisément - et plus communément - en pourcentages (auquel cas S est le produit des profils saisonnier et de rotation des jours). La variante additive produirait des interpolations Γ^* à saisonnalité mensuelle infra-trimestrielle négligeable dans les cas où les chiffres financiers sont en millions (disons). De telles interpolations seraient généralement d'une précision insuffisante pour produire un trimestrialisation satisfaisante des trimestres financiers F.

Il y aurait alors trois options. L'une consisterait à multiplier S par des facteurs de calibration évoluant de manière graduelle de mois en mois et ensuite appliquer la variante additive au S calibré. La seconde option consisterait à adapter la variante proportionnelle de Denton (1971) à la trimestrialisation. Une telle variante résoudrait bien le problème de calibration, en gardant Γ^* proportionnel à S; cependant les poids W et P (de (3.8) et (3.11) dépendraient alors des données. La troisième option est d'adopter la variante logarithmique exposée ici.

Pour les séries de stock, la variante logarithmique consiste simplement à appliquer la variante additive aux logarithmes des trimestres financiers, ln F, et du profil saisonnier, ln S; et à poser les interpolations désirées égales à l'anti-logarithme des estimations obtenues. Ainsi, la solution (3.8) devient:

$$\ln \Gamma^* = \ln S + W\ [\ \ln F - B\ \ \ln S\ ], \qquad \Gamma^* = \exp(\ln \Gamma^*) \tag{4.1}$$

où les poids W sont ceux de (3.8) (avec $V_\epsilon=0$). Puisque la variante additive préserve le changement d'un mois à l'autre de S, Γ^* de (4.1) préserve le taux de croissance d'un mois à l'autre.

$$S = \Gamma + e, \quad E(e)=0, \quad E(e\,e') = V_e \quad ; \tag{3.1a}$$

$$F = B\,\Gamma + \epsilon, \quad E(\epsilon)=0, \quad E(\epsilon\,\epsilon') = V_\epsilon = \sigma_\epsilon^2\, I, \quad \epsilon\rightarrow 0, \quad \sigma_\epsilon^2\rightarrow 0. \tag{3.1b}$$

Dans le contexte de la trimestrialisation, le vecteur S de dimensions T par 1 représente une variable auxiliaire mensuelle. Dans ce travail, et sans perte de généralité, la variable auxiliaire S prend la forme d'un profil saisonnier ou d'un profil saisonnier plus un profil de rotation des jours (Young, 1965). Ce profil est valable pour tous les répondants, au niveau d'agrégation auquel s'effectue la trimestrialisation. Le vecteur F de dimensions M par 1 contient les jalons de trimestres financiers, c-à-d les données à trimestrialiser. Le vecteur Γ contient les T valeurs mensuelles inconnues à estimer.

La matrice B de dimensions M par T est un opérateur de sommes trimestrielles financières. Par exemple, dans le cas d'une série de flux à trimestres financiers couvrant de février à avril, mai à juillet, etc, la matrice B serait:

$$\underset{M \text{ by } T}{B} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & & & & & & \end{bmatrix}, \tag{3.2}$$

(Pour les séries de stock, des 0 remplacent les deux premiers 1 de chaque rangée.) Par conséquent, l'équation (3.1b) spécifie que les sommes financières des interpolations désirées Γ sont égales aux données disponibles de trimestres financiers (sauf pour une erreur infinitésimalement petite dont la présence deviendra bientôt évidente).

Enfin la matrice de covariance V_e des erreurs $e = [e_t,\ t=1,\ldots,T]$ est telle que e_t change le moins possible du mois t au mois t+1:

$$\underset{T \text{ by } T}{V_e} = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix} \sigma_u^2 \;/\; (1-\rho^2). \tag{3.3}$$

où ρ est inférieur mais très voisin de 1 (ex. 0.999999) et où σ_u^2 est en pratique la variance du changement dans S (c-à-d la variance de $(S_t - S_{t-1})$). En d'autres mots, cette matrice spécifie une extrême autocorrélation des erreurs au décalage 1. (Détails dans Cholette et Dagum, 1989; Cholette et Baldwin, 1989). L'effet de V_e dans (3.1a) est de garder les valeurs interpolées estimées Γ^* le plus parallèle possible au profil saisonnier choisi S. Le degré de parallélisme atteint dépend des jalons financiers dans (3.1b).

3.2 La solution

Le modèle (3.1) peut s'écrire

$$Y = X\,\Gamma + U, \quad E(U)=0, \quad E(U\,U')=V, \tag{3.4}$$

où:

$$Y' = [\,S' \;\; F'\,], \quad X' = [\,I \;\; B'\,], \quad U' = [\,e' \;\; \epsilon'\,], \quad V = \begin{bmatrix} V_e & 0 \\ 0 & V_\epsilon \end{bmatrix}.$$

La solution de (3.4) par les moindres carrés généralisés est:

$$\Gamma^* = (X'V^{-1}X)^{-1}\; X'V^{-1}\; Y = [V_e^{-1} + B'V_\epsilon^{-1}\,B\,]^{-1}\; [V_e^{-1}\,S + B'V_\epsilon^{-1}\;F\,], \tag{3.5}$$

$$\text{var } \Gamma^* = (X'V^{-1}X)^{-1} = [V_e^{-1} + B'V_\epsilon^{-1}\,B\,]^{-1}, \tag{3.6}$$

où V_e^{-1} est connu algébriquement:

$$V_e^{-1} = \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 & \ldots \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & \ldots \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \;/\; \sigma_u^2. \tag{3.7}$$

À l'aide d'identités d'algèbre matricielle, la solution (3.5) peut s'exprimer:

$$\Gamma^* = S + V_e\,B'[B\,V_e\,B' + V_\epsilon]^{-1}\;[\,F - B\,S\,] = S + W\,[\,F - B\,S\,] = S + W\,R. \tag{3.8}$$

trimestres civils de chaque année dans l'exemple. Ces erreurs constituent en outre un biais: chaque année les premier et troisième trimestres se trouvent *systématiquement* sur-évalués tandis que les deuxième et quatrième sont sous-évalués. En présence de saisonnalité surtout, l'assimilation des trimestres financiers aux trimestres civils, c'est-à-dire la négligence du problème de trimestrialisation, cause erreur et biais dans la série trimestrielle résultante.

Lors d'une trimestrialisation véritable, les valeurs mensuelles de la figure 1 sont évidement inconnues. La stratégie proposée dans ce travail consiste en deux étapes:

(1) interpoler les chiffres mensuels inconnus, à partir des chiffres trimestriels financiers et d'une variable auxiliaire, généralement sous forme de profil saisonnier; et

(2) poser les estimations trimestrielles civiles égales aux sommes proprement trimestrielles des interpolations mensuelles.

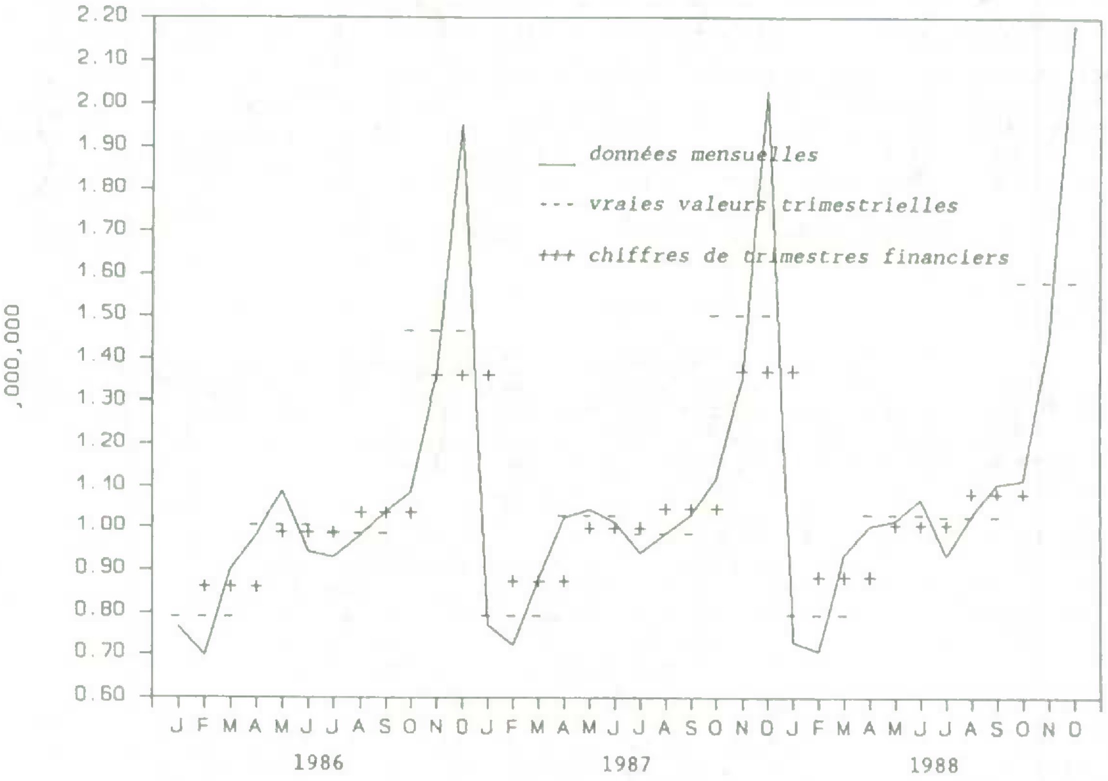


Figure 1: Différences entre les valeurs trimestrielles civiles (---) et les valeurs trimestrielles financières (+++) des ventes mensuelles canadiennes des Grands magasins

3. VARIANTE ADDITIVE DE LA TRIMESTRIALISATION

La présente section expose les méthodes d'étalonnage du type Denton comme une régression linéaire et l'adapte aux fins de la trimestrialisation. Les instituts de statistique utilisent normalement l'étalonnage lorsque, pour une même variable socio-économique, ils ont des mesures infra-annuelles (disons) ainsi que des mesures annuelles, obtenues d'une autre source plus fiable et considérées comme jalons. Dans pareils cas, les sommes annuelles de la série infra-annuelles diffèrent généralement des jalons annuels correspondants. Pour les méthodes du type Denton, l'étalonnage consiste alors à ajuster la série infra-annuelle de sorte à ce que (a) les totaux annuels de la série étalonnée concordent avec les jalons et que (b) la étalonnée soit le plus parallèle possible à la série infra-annuelle originale. La méthode de trimestrialisation proposée consiste essentiellement

(1) à étalonner un profil saisonnier aux données disponibles de trimestres financiers, considérés comme jalons; et

(2) à prendre les sommes sur trimestres civils de la série ajustée aux jalons.

La première étape fournit les estimations des valeurs mensuelles inconnues, c.-à-d. les interpolations; et la seconde, les valeurs proprement trimestrielles désirées.

3.1 Le modèle

Selon Cholette et Dagum (1989), la méthode de Denton peut se voir comme une régression, comprenant deux équations:

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

LA TRIMESTRIALISATION DES CHIFFRES DE TRIMESTRES FINANCIERS

P.A. Cholette[1]

RÉSUMÉ

Beaucoup d'enquêtes trimestrielles effectuées par les instituts de statistique reflètent en fait les trimestres financiers des répondants, couvrant par exemple les mois de février à avril, de mai à juillet, etc. Ce travail propose une méthode pour transformer de telles données en estimations proprement trimestrielles, couvrant de janvier à mars, avril à juin, etc.

La méthode est essentiellement une adaptation de la méthode d'étalonnage de Denton (1971): Il s'agit d'étalonner un profil saisonnier de sorte à le rendre compatible avec les jalons trimestriels financiers. Les estimations trimestrielles sont ensuite données par les sommes trimestrielles appropriées des valeurs mensuelles "étalonnées". La méthode de Denton est présentée à neuf dans le cadre familier de la régression.

MOTS CLEFS: Étalonnage, Interpolation, Trimestres financiers, Années financières, Désagrégation temporelle.

1. INTRODUCTION

Toutes les enquêtes trimestrielles effectuées par Statistique Canada portent en fait sur les trimestres financiers des répondants. Ces trimestres couvrent trois mois consécutifs quelconques, par exemple février à avril, mai à juillet, etc.; ou bien mars à mai, juin à août, etc. Il arrive même que les «mois» en question ne se terminent pas le dernier jour du mois. Dans certains cas évidemment, les trimestres financiers coïncident avec les trimestres civils, portant sur janvier à mars, avril à juin, etc.

Une pratique à l'égard des chiffres de trimestres financiers consiste à les assimiler au trimestre civil qui chevauche le plus. Par exemple, si les répondants d'une enquête ont l'un ou l'autre des trimestres financiers suivants, décembre à février, janvier à mars et février à avril, leurs réponses sont toutes assimilées au premier trimestre. Le total «trimestriel» de ces réponses reflète donc implicitement cinq mois (décembre à avril) au lieu du premier trimestre. En régime saisonnier surtout, pareils chiffres trimestriels sont évidemment trompeurs.

Ce travail propose une méthode pour trimestrialiser les chiffres de trimestres financiers, c'est-à-dire pour les transformer en valeurs trimestrielles civiles. On suppose (1) que les répondants dans l'enquête ont des trimestres financiers communs, ou qu'au moins la trimestrialisation s'effectue à un niveau d'agrégation où cela se vérifie; et (2) que les trimestres financiers se terminent en fin de mois. La section 2 illustre le problème de la trimestrialisation sous ces deux hypothèses simplificatrices.

La section 3 présente la variante additive de la méthode de trimestrialisation proposée, qui constitue en fait une adaptation des méthodes d'étalonnage de type Denton (ex. Denton, 1971; Helfand, Monsour et Trager, 1977). (L'étalonnage consiste à ajuster une série infra-annuelle à des valeurs annuelles obtenues d'une autre source plus fiable.) La section 4 présente une variante logarithmique de la méthode proposée. La section 5 suggère une mise en oeuvre économique des deux variantes et examine la question des révisions des estimations. La section 6 met à l'épreuve la méthode en l'appliquant à dix séries canadiennes de commerce au détail.

2. LE PROBLÈME DE LA TRIMESTRIALISATION

Le problème de la trimestrialisation des trimestres financiers se décrit aisément à l'aide d'une illustration. La figure 1 montre trois années de ventes mensuelles des grands magasins canadiens. La figure montre également les valeurs de trimestres civils, représentées par leur moyenne sur les trois mois couverts (c.-à-d. divisées par 3), ainsi que celles de trimestres financiers (aussi divisées par 3), couvrant les mois de février à avril, mai à juillet, etc. Il appert que l'assimilation des valeurs trimestrielles financières au trimestre le plus voisin (qui chevauche le plus) crée des erreurs d'«estimation» substantielles, spécialement pour les premier et dernier

[1] P.A. Cholette, Division de la recherche et de l'analyse des séries chronologiques, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6

Le processus itératif décrit ci-dessus cause de nombreux problèmes de convergence, que Draper et Smith (1981) exposent avec précision tout en proposant des solutions. En ce qui concerne la méthode décrite en 3.4, on peut chercher à résoudre ces difficultés en exploitant la structure du modèle. On peut notamment procéder de la façon suivante: i) pour une valeur fixe, définir l'estimateur par les moindres carrés pondérés linéaires de θ_t comme une fonction de α, par exemple $\theta_t(\alpha)$, et ii) utiliser $\hat{\theta}_t(\alpha)$ au lieu de θ_t dans le modèle d'étalonnage et appliquer les MCP non linéaires pour obtenir une estimation de α, c'est-à-dire $\hat{\alpha}$. De cette façon, on fait passer la dimension de l'algorithme de Gauss-Newton de n+1 à 1.

L'expression définie ci-dessus pour $\underline{\gamma}_{j+1}$ suppose que les valeurs annuelles sont observées avec une erreur, de sorte que $\underline{\Sigma}_u$ est non singulière. Toutefois, cette matrice de covariances sera singulière lorsque les valeurs annuelles viendront d'un recensement. Dans ces conditions, on obtiendra la solution en remplaçant $\underline{L}_j$ par la g-inverse de semi-norme $\underline{\Sigma}_u$ minimum. Autrement dit, $\underline{\Sigma}_u + \underline{J}_j \underline{J}_j'$ est substituée à $\underline{\Sigma}_u$ dans l'équation de $\underline{L}_j$ (voir Rao et Mitra, 1971).

2. MATRICE DES VARIANCES-COVARIANCES DES ESTIMATIONS

En supposant que l'algorithme de Gauss-Newton converge vers les valeurs estimées $\hat{\gamma} = \gamma_j$, après j itérations, une approximation de la matrice des covariances est alors donnée par Var $(\hat{\underline{\gamma}}) = \underline{L}_j' \underline{\Sigma}_u \underline{L}_j$.

3. SOURCES ADDITIONNELLES

Draper, N.R. et Smith, H. (1981), *Applied Regression Analysis*, seconde édition, New York, Wiley.

Rao, C.R. et Mitra, S.K. (1971), *Generalized Inverse of Matrices and Its Applications*, New York, Wiley.

ANNEXE

1. ALGORITHME DE GAUSS-NEWTON

Les modèles pour estimations annuelles et infra-annuelles définis dans les sous- sections 3.3 et 3.4 peuvent se résumer en un modèle défini comme suit:

$$Y_s = f(\underline{X}_s, \underline{\gamma}) + u_s \qquad \text{for } s = 1, \ldots, n + m$$

où: Y_s représente la réponse infra-annuelle lorsque $s = 1, \ldots, n$ et la réponse annuelle lorsque $s = n + 1, \ldots, n + m$,

$\underline{X}_s$ $= (X_{1s}, \ldots, X_{n+m,s})'$ est un vecteur de variables dichotomiques définies ainsi:

$$X_{ks} = \begin{matrix} 1 \text{ si } s = k \\ 0 \text{ si } s = k \end{matrix} ;$$

$\underline{\gamma}$ $= \gamma_1, \ldots, \gamma_p)'$ est le vecteur des paramètres à estimer dans le modèle annuel et infra-annuel combiné,

u_s est l'erreur d'échantillonnage infra-annuelle lorsque $s = 1, \ldots, n$ et l'erreur d'échantillonnage annuelle lorsque $s = n + 1, \ldots, n + m$;

$f(\underline{X}_s, \underline{\gamma})$ est égale à $\sum_{k=1}^{g} g_k(\underline{\gamma}) X_{ks}$, où $g_k(\gamma)$ représente le modèle infra-annuel lorsque $k = 1, \ldots, n$ et le modèle annuel lorsque $k = n + 1, \ldots, n + m$.

Par exemple, en ce qui concerne la méthode décrite en 3.4, nous avons $\underline{\gamma} = (\alpha, \theta_1, \ldots, \theta_n)'$ et

$$g_k(\underline{\gamma}) = \begin{matrix} \alpha\theta_k & \text{si } k = 1, \ldots, n \\ \sum_{t \epsilon k} \theta_t & \text{si } k = n + 1, \ldots, n + m. \end{matrix}$$

Les modèles pour données infra-annuelles définis dans les sous-sections 3.3 et 3.4 sont tous deux non linéaires par rapport aux paramètres. Dans les circonstances, nous pouvons estimer les paramètres par la méthode de linéarisation, qui consiste à faire une approximation du modèle non linéaire à l'aide d'un modèle linéaire de la forme $Y_s - f_s^0 = \sum_{i=1}^{p} \beta_i^0 J_{is}^0 + u_s$

où: $f_s^0 = f(\underline{X}_s, \underline{\gamma}_0)$, $\beta_i^0 = \gamma_i - \gamma_{i0}$, $\underline{\gamma}_0 = (\gamma_{10}, \ldots, \gamma_{p0})'$, et $J_{is}^0 = \left. \dfrac{\partial f(\underline{X}_s, \underline{\gamma})}{\partial \gamma_i} \right|_{\underline{\gamma} = \underline{\gamma}_0}$

sont les estimations initiales voisines des valeurs réelles. Nous avons établi ces estimations au moyen de la méthode d'étalonnage de Denton (voir sous-section 3.1).

Il y a moyen d'améliorer les estimations initiales en utilisant les moindres carrés linéaires dans des itérations successives, ce qui donne l'équation matricielle suivante:

$$\underline{\gamma}_{j+1} = \underline{\gamma}_j + \underline{L}_j' (\underline{Y} - \underline{f}^j)$$

where

$$\underline{\gamma}_j = (\gamma_{ij}, \ldots, \gamma_{pj})' \qquad \underline{f}^j = (f_1^j, \ldots, f_{n+m}^j)' \qquad \underline{J}_j = \{J_{is}^j\}_{(n+m) \times p} \qquad \underline{u} = (u_1, \ldots, u_{n+m})'$$

$$\underline{Y} = (Y_1, \ldots, Y_{n+m})' \qquad \underline{L}_j = \underline{J}_j' \underline{\Sigma}_u^{-1} \underline{J}_j^{\;-1} \underline{J}_j' \underline{\Sigma}_u^{-1\;\;'} \qquad \underline{\Sigma}_u = E(\underline{uu}')$$

A cause des erreurs d'arrondissement, la matrice $\underline{J}_j' \underline{\Sigma}_u^{-1} \underline{J}_j$ peut paraître singulière et, parconséquent, non inversible. Cela s'explique par le fait que certains éléments de $\underline{J}_j$ diffèrent largement les uns des autres. Pour résoudre cette difficulté, il suffit de diviser la série de données infra-annuelles et la série de données annuelles par la moyenne des niveaux infra-annuels avant d'appliquer l'algorithme d'itération. Une fois la convergence réalisée, on fait l'opération inverse (multiplication) pour obtenir les estimations infra-annuelles étalonnées.

Cela revient à supposer que l'étalonnage est une opération annuelle. On devrait donc pouvoir tirer de cette opération des facteurs d'étalonnage estimés qui serviront à un étalonnage préliminaire.

Il y a deux façons de produire des facteurs d'étalonnage préliminaires:

1) Reprendre le facteur qui a été calculé pour la dernière période infra- annuelle qui a fait l'objet d'un étalonnage soit:

a) en appliquant l'étalonnage jusqu'à la dernière période infra- annuelle pour laquelle il existe des données annuelles, ou

b) en appliquant l'étalonnage jusqu'à la dernière période infra- annuelle pour laquelle il existe des données infra-annuelles.

2) A l'aide d'un modèle, extrapoler la série infra-annuelle jusqu'à la période infra-annuelle où doit se faire le prochain étalonnage. Ensuite, procéder à l'étalonnage en se servant de la série extrapolée pour obtenir les facteurs préliminaires. Laniel (1986) propose des modèles simples pour exécuter ce genre d'extrapolation. Il faut s'assurer que ces modèles sont suffisamment robustes pour de grandes enquêtes de manière à ne pas produire de facteurs préliminaires qui soient moins fiables qu'une méthode qui ne fait que reprendre le dernier facteur d'étalonnage calculé.

Ces deux méthodes méritent d'être analysées plus en profondeur. On pourrait notamment examiner les corrections apportées aux données étalonnées, depuis les données préliminaires jusqu'aux données finales.

6. CONCLUSION

Dans cet article, nous avons vu comment améliorer des estimations d'enquête infra-annuelles au moyen d'estimations d'enquête annuelles. Nous avons présenté un moyen simple et inédit d'étalonner une série chronologique. La méthode en question peut être appliquée par ordinateur en mode automatique. Elle se distingue avantageusement des méthodes plus classiques par sa base statistique. Elle permet en effet de déterminer des régions de confiance et de tester la validité de l'ajustement du modèle d'étalonnage. Par la même occasion, nous avons abordé certaines questions liées à l'utilisation de cette méthode. Deux grandes questions pratiques ont retenu notre attention: l'étalonnage d'un tableau de données chronologiques et l'étalonnage préliminaire. Nous avons proposé des façons d'aborder ces deux questions mais les recherches doivent se poursuivre.

7. BIBLIOGRAPHIE

Box, G.E.P. et Jenkins, G.M. (1976), "Time Series Analysis, Forecasting and Control", Holden-Day.

Cholette, P.A. (1988a), "Benchmarking and Interpolation of Time Series", Statistics Canada, Working Pager No. TSRA-87-014E.

Cholette, P.A. (1988b), "Benchmarking Systems of Socio-Economic Time Series", Statisties Canada Working Paper No. TSRA-88-017E.

Cholette, P.A. et Dagum, E.B. (1989), "Benchmarking Socio-Economic Time Series Data: A Unified Approach", Working Paper No. TSRA-89-006-E, Statistics Canada.

Deming, W.E. et Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known", Ann. Math. Statist.

Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", Journal of the American Statistical Association, Vol. 66, No. 333, pp. 99-102.

Hillmer, S.C. et Trabelsi, A. (1987), "Benchmarking of Economic Time Series", Journal of the American Statistical Association, Vol. 82, pp. 1604-1071.

Laniel, N. (1986), "Adjustment of Economic Production Sub-annual Series", Business Survey Methods, Statistics Canada, Working Paper No. BSMD-86-006E.

Laniel N. et Fyfe, K. (1989), "Benchmarking of Economic Time Series", Business Survey Methods, Statistics Canada, Business Survey Redesign Project Working Paper.

Monsour, N.J. et Trager, M.L. (1979), "Revision and Benchmarking of Business Time Series", Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Quenneville, B. et Srinath, K.P. (1984), "Estimation of Variances of Averages Based on Overlapping Samples in Repeated Surveys", Proceedings of the Survey Research Methods Section, American Statistical Association.

étalonnées, θ_t, ont un très petit nombre de degrés de liberté, $m - 1$ (nombre d'observations annuelles moins un), par rapport au nombre d'observations, $n + m$. De plus, on peut s'attendre que les estimations étalonnées affichent la même tendance chronologique que les données infra-annuelles.

A ce stade-ci, le calcul des covariances d'échantillonnage de deux estimations de niveau se rapportant à deux périodes différentes est une question d'ordre pratique. Devrait-on modéliser ces covariances ou les calculer directement pour toutes les paires de périodes au moyen d'un logiciel d'estimation? Du point de vue théorique, il est préférable de les calculer directement puisque la suite des erreurs d'échantillonnage est en soi un processus stochastique non stationnaire, mais il n'est pas évident que nous puissions procéder ainsi. En revanche, nous ne connaissons pas de modèle dont la validité ait été vérifiée jusqu'à maintenant. Certains auteurs ont expérimenté arbitrairement un modèle stationnaire AR(1) (voir Hillmer et Trabelsi, 1987) mais celui-ci ne semble pas valide a priori. Quenneville et Srinath (1984) ont abordé la question sous un angle légèrement différent en modélisant la corrélation d'échantillonnage entre les périodes à l'aide de la structure d'autocorrélation d'un processus AR(1). Toutefois, cette méthode n'a pas donné de résultats concluants. Il faut donc continuer de chercher une solution au problème du calcul des covariances d'échantillonnage.

4. ÉTALONNAGE D'UN TABLEAU DE DONNÉES CHRONOLOGIQUES

La plupart des enquêtes économiques infra-annuelles produisent des séries d'estimations pour un certain nombre d'activités industrielles qui se déroulent dans un certain nombre de régions géographiques. Ces estimations sont publiées infra-annuellement dans des tableaux dont il faut étalonner les diverses composantes (fréquences par case, totaux marginaux et total général).

Si les séries de fréquences par case, les séries de totaux marginaux et la série de totaux généraux sont toutes étalonnées de façon indépendante, nous obtiendrons une série d'estimations infra-annuelles où la somme des fréquences par case ne correspondra pas au total marginal et où la somme des totaux marginaux ne correspondra pas au total général. Autrement dit, nous obtiendrons une série de tableaux incohérents. Pour éviter cela, nous pouvons envisager un certain nombre de solutions. La première qui nous vient à l'esprit est la méthode élémentaire suivante: premièrement, étalonnage de chaque série de fréquences par case; deuxièmement, addition des fréquences étalonnées de manière à obtenir des totaux marginaux étalonnés et un total général étalonné. L'inconvénient de cette méthode est que la tendance chronologique des valeurs étalonnées risque d'être plus irrégulière que celle que l'on observerait par suite d'un étalonnage direct (problème courant en désaisonnalisation). Si tel est le cas, il sera préférable d'utiliser la méthode suivante:

i) Premièrement, étalonner la série de totaux généraux.

ii) Ensuite, étalonner de façon indépendante chaque série de totaux marginaux puis pour chaque période infra-annuelle, redresser chacun des totaux marginaux étalonnés par un facteur constant de manière que leur somme corresponde au total général étalonné.

iii) Finalement, étalonner chaque série de fréquences par case puis pour chaque période infra-annuelle, redresser chacune des fréquences étalonnées par la méthode itérative du quotient (aussi appelée ajustement proportionnel itératif, voir Deming et Stephan, 1940) de manière que leur somme corresponde à la valeur redressée du total marginal étalonné correspondant.

Cette méthode suppose que la série de totaux généraux est la plus importante de toutes en ce qui concerne le maintien des tendances d'un mois à l'autre, que la série des totaux marginaux est la seconde en importance et que celle des fréquences par case vient au troisième rang. L'inconvénient est que les tendances d'un mois à l'autre des fréquences par case peuvent être fortement perturbées. C'est ce qui a été observé dans un petit nombre de cas (voir Laniel et Fyfe, 1989).

On pourrait aussi envisager d'étalonner simultanément les séries de fréquences par case et de totaux marginaux et la série de totaux généraux. Cependant, le problème pourrait alors prendre des proportions gigantesques si on pense au nombre de paramètres à estimer, et pourrait même être difficile à traiter sur ordinateur. Cholette (1988b) s'est penché sur la question lorsqu'il a étudié l'étalonnage des séries par la méthode de Denton.

Il est nécessaire de pousser plus loin la recherche sur ces trois méthodes avant de déterminer laquelle s'appliquerait le mieux au problème décrit dans la section 2.

5. ÉTALONNAGE PRÉLIMINAIRE

On procède à un étalonnage préliminaire pour empêcher qu'il y ait discontinuité entre les périodes infra-annuelles pour lesquelles il existe des données annuelles et celles pour lesquelles il n'en existe pas. Il peut se produire des discontinuités parce que les données relatives à une année civile particulière sont diffusées environ dix-huit mois après la fin de cette année civile. Il existe donc deux genres de périodes infra-annuelles pour lesquelles nous n'avons pas toujours de données annuelles: les périodes pour lesquelles il existe des estimations infra-annuelles et celles pour lesquelles nous n'aurons des estimations infra-annuelles qu'au prochain étalonnage.

$\{\varepsilon_t\}$ est une suite d'erreurs d'échantillonnage corrélées infra-annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0}, \underline{\Sigma}_\varepsilon)$,

$\{z_T\}$ est une suite d'estimations annuelles sans biais,

$\{\alpha_T\}$ est une suite d'erreurs d'échantillonnage corrélées annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0}, \underline{\Sigma}_a)$.

On obtient les estimations étalonnées en appliquant les moindres carrés aux modèles ci-dessus. L'algorithme de Gauss-Newton utilisé à cette fin est décrit en annexe. Cette description est suivie du calcul de la matrice des covariances des estimations étalonnées.

On peut utiliser cette méthode lorsque les données repères proviennent d'un recensement ou d'une enquête annuelle avec échantillons chevauchants et lorsque les estimations de niveau infra-annuelles sont biaisées, à la condition que le biais relatif soit fixe. En pratique, l'hypothèse d'un biais relatif constant se vérifiera lorsque les opérations de mise à jour de la base de sondage se feront à un rythme régulier, c'est-à-dire lorsque la proportion d'unités absentes de la base de sondage se sera stabilisée avec les années. De plus, on suppose que les entreprises non recensées ont le même comportement que celles qui figurent dans la base de sondage. Ces hypoth`ses se vérifient lorsque la méthode d'étalonnage est appliquée à un petit nombre d'années à la fois.

La méthode décrite ci-dessus pose toutefois un problème technique. En effet, on ne peut calculer directement la matrice des variances-covariances d'échantillonnage des tendances et on doit donc recourir à une approximation. Après avoir utilisé l'approximation de Taylor du premier degré, on a constaté que dans certains cas, les variances et covariances d'échantillonnage étaient nulles ou négatives alors qu'elles devraient être positives.

3.4 Modèle pour niveaux

La méthode ci-dessous équivaut à la précédente et a été élaborée dans le but de calculer plus facilement la matrice des variances-covariances d'échantillonnage des estimations infra-annuelles. Elle suppose que les données infra-annuelles sont expliquées par le modèle:

$$y_t = \alpha\theta_t + \varepsilon_t \qquad t = 1, 2, \ldots, n$$

et les données annuelles, par le modèle:

$$z_T = \sum_{t\varepsilon T} \theta_t + a_T \qquad T = 1, 2, \ldots, m$$

où:

$\{y_t\}$ est une suite d'estimations biaisées des niveaux infra-annuels,

α est un paramètre fixe qui tient compte du biais relatif constant,

$\{\theta_t\}$ est une suite de paramètres infra-annuels fixes (valeurs réelles des niveaux),

$\{\varepsilon_t\}$ est une suite d'erreurs d'échantillonnage corrélées infra-annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0}, \underline{\Sigma}_\varepsilon)$,

$\{z_T\}$ est une suite d'estimations annuelles non biaisées,

$\{a_T\}$ est une suite d'erreurs d'échantillonnage corrélées annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0}, \underline{\Sigma}_a)$.

On obtient les estimations étalonnées en appliquant les moindres carrés aux modèles ci-dessous. L'algorithme pertinent est le même que celui utilisé pour la méthode décrite en 3.3.

On peut recourir à la méthode ci-dessus lorsque les données annuelles proviennent soit d'un recensement ou d'une enquête avec échantillons chevauchants et lorsque les estimations de niveau infra-annuelles sont biaisées, à la condition que le biais relatif soit fixe dans le temps.

3.5 Discussion

De toutes les méthodes exposées ci-dessus, la dernière est celle qui se prête le mieux à l'étalonnage d'une série chronologique. Elle offre une base statistique qui permet de calculer des régions de confiance et de tester la validité de l'ajustement du modèle étalonné. Le choix du test doit être exercé avec soin puisque les estimations vérifient avec des séries économiques. Bien que cette méthode permette de composer avec des données infra-annuelles entachées d'une erreur systématique, elle ne tient aucunement compte des variances et des covariances d'échantillonnage des données annuelles et infra-annuelles, ce qui la rend statistiquement inefficace.

3.2 Méthode de Hillmer et Trabelsi

En 1987, Hillmer et Trabelsi ont proposé une méthode d'étalonnage fondée sur les modèles ARMMI de Box et Jenkins (1976). Ils ont supposé que les données infra- annuelles étaient expliquées par le modèle:

$$y_t = \theta_t + \varepsilon_t \qquad t = 1,2,\ldots,n$$

et les données annuelles, par le modèle:

$$z_T = \sum_{t \varepsilon T} \theta_T + a_t \qquad T = 1,\ 2,\ \ldots,\ m$$

où:

$\{\theta_t\}$ est une suite de paramètres infra-annuels stochastiques (valeurs réelles des niveaux) qui suivent un modèle ARMMI,

$\{y_t\}$ est une suite d'estimations non biaisées des paramètres infra-annuels,

$\{\varepsilon_t\}$ est une suite d'erreurs d'échantillonnage corrélées infra-annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0},\ \underline{\Sigma}_\varepsilon)$

$\{z_T\}$ est une suite d'estimations annuelles non biaisées,

$\{\alpha_T\}$ est une suite d'erreurs d'échantillonnage corrélées annuelles avec un vecteur de moyennes et une matrice de covariances $(\underline{0},\ \underline{\Sigma}_a)$

Se servant de ces modèles, Hillmer et Trabelsi obtiennent des estimations infra-annuelles étalonnées en appliquant les moindres carrés stochastiques, c'est-à- dire en minimisant l'erreur quadratique moyenne, $E(\hat{\theta}_t - \theta_t)^2$. Dans la terminologie de l'analyse des séries chronologiques, cette méthode est aussi appelée extraction de signal et Hillmer et Trabelsi en font l'illustration dans leur article.

Avec cette méthode, les données annuelles proviennent soit d'un recensement ou d'une enquête avec échantillons chevauchants. De plus, les variances et les covariances d'échantillonnage des estimations de niveau infra-annuelles sont prises en considération. Malheureusement, cette méthode ne tient pas compte des erreurs systématiques que peuvent renfermer les données infra-annuelles des enquêtes économiques. En outre, comme il s'agit de modèles ARMMI, il serait trop coûteux d'utiliser cette méthode pour de grandes enquêtes d'où sont tirées des centaines de séries de données. Il serait donc plus profitable de n'utiliser cette méthode que pour un petit nombre d'indicateurs économiques majeurs. Si les modèles ARMMI sont mal spécifiés, on risque par ailleurs d'obtenir un lissage excessif des données.

Cholette et Dagum (1989) ont perfectionné la méthode de Hillmer et Trabelsi en utilisant un modèle "d'intervention" au lieu d'un modèle ARMMI. Cela permet de modéliser les effets systématiques contenus dans la série chronologique. Selon les auteurs toutefois, cette méthode présente les mêmes lacunes que la méthode de Hillmer et Trabelsi.

3.3 Modèle pour tendances

La méthode suivante a été élaborée dans le but de satisfaire aux conditions d'étalonnage des enquêtes économiques. Elle repose sur l'hypothèse que les données infra-annuelles sont expliquées par le modèle:

$$\frac{y_t}{y_{t-1}} = \frac{\theta_t}{\theta_{t-1}} + \varepsilon_t \qquad t = 1,\ 2,\ \ldots,\ n$$

et les données annuelles, par le modèle:

$$z_T = \sum_{t \varepsilon T} \theta_t + a_T \qquad T = 1,\ 2,\ \ldots,\ m$$

où:

$\{y_t/y_{t-1}\}$ est une suite d'estimations (quasi) sans biais des tendances infra-annuelles,

$\{\theta_t/\theta_{t-1}\}$ est une suite de tendances des paramètres infra-annuels fixes (valeurs réelles), entreprises qui n'ont pas d'employés et la plupart des nouvelles entreprises. En outre, les données annuelles proviennent habituellement de grands échantillons chevauchants, ce qui fait que l'on peut leur associer des erreurs d'échantillonnage.

Lorsqu'on procède à un étalonnage, il faut se rappeler que les résultats des enquêtes annuelles sont produits environ deux ans après que les enquêtes ont été réalisées. Par exemple, les données annuelles pour 1988 ne seront pas diffusées avant 1990 tandis que les données infra-annuelles sont diffusées habituellement quelques mois suivant la période à laquelle elles s'appliquent. Ainsi, lorsque vient le moment d'étalonner les données infra-annuelles, il y a des périodes infra-annuelles pour lesquelles nous n'avons pas de repères annuels.

Une méthode d'étalonnage doit présenter un certain nombre de caractéristiques pour pouvoir être appliquée à des estimations de grandes enquêtes. Premièrement, elle doit être suffisamment simple pour pouvoir être utilisée en mode automatique et nécessiter le moins possible l'intervention du statisticien. Deuxièmement, elle doit pouvoir produire des facteurs d'étalonnage préliminaires pour les mois pour lesquels il n'existe pas encore de données repères. Cette caractéristique permet de faire de l'étalonnage à mesure que sont produites les données infra-annuelles, autrement il se créerait des discontinuités.

La méthode d'étalonnage utilisée doit pouvoir produire de meilleures estimations de niveau et de meilleures estimations de tendance d'une année à l'autre tant pour les flux (c.-à-d. données infra-annuelles qui ont trait à un intervalle de temps, par exemple: ventes) que pour les stocks (c.-à-d. données infra-annuelles qui ont trait à une date précise, par exemple: stock de marchandises). Elle devrait aussi pouvoir maintenir la concordance des totaux généraux, des totaux marginaux et des estimations par case pour les données étalonnées provenant d'une série de tableaux.

3. ÉTALONNAGE D'UNE SÉRIE CHRONOLOGIQUE

Nous allons exposer ci-dessous quatre méthodes qui peuvent servir à étalonner une série de données infra-annuelles sur les flux ou les stocks. Dans chaque cas, nous donnons une interprétation statistique de la méthode, décrivons brièvement les hypothèses sous-jacentes et faisons une évaluation qualitative du bien-fondé de la méthode à l'égard du problème exposé dans la section précédente.

3.1 Méthode de Denton

Dans son article de 1971, Denton a proposé des méthodes d'étalonnage fondées sur la minimisation quadratique. Chacune de ces méthodes correspond à une fonction de perte particulière et l'une d'elles peut être utilisée pour résoudre le problème d'étalonnage exposé dans la section 2 à la condition que certaines hypothèses concernant les données soient vérifiées. La méthode en question utilise une fonction de perte qui s'exprime sous forme d'écarts proportionnels entre la série originale et la série étalonnée. Nous pouvons exprimer cette méthode en termes statistiques en supposant tout d'abord que les données infra- annuelles sont expliquées par le modèle:

$$\frac{\theta_t}{y_t} = \frac{\theta_{t-1}}{y_{t-1}} + \varepsilon_t \qquad t = 1, 2, \ldots, n$$

et les données annuelles, par le modèle:

$$z_T = \sum_{t \varepsilon T} \theta_t \qquad T = 1, 2, \ldots, m$$

où:

- t indique la période infra-annuelle
- T indique la période annuelle
- $\{y_t\}$ est une suite d'estimations biaisées des paramètres infra-annuels (niveaux),
- $\{\theta_t\}$ est une suite de paramètres infra-annuels fixes (valeurs réelles des niveaux),
- $\{\varepsilon_t\}$ est une suite d'erreurs non corrélées et identiquement distribuées avec un vecteur de moyennes et une matrice de covariances $(0, \sigma^2 \underline{I})$
- $\{z_t\}$ est une suite de valeurs repères annuelles tirées d'un recensement.

Pour obtenir les estimations étalonnées, on applique les moindres carrés au second modèle défini ci-dessus.

Il convient de souligner que la méthode de Denton suppose que θ_t / y_t suit une marche aléatoire et que les données annuelles proviennent d'un recensement. Malheureusement, il est peu probable que ces hypothèses se

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ÉTALONNAGE DES SÉRIES ÉCONOMIQUES

N. Laniel et K. Fyfe[1]

RÉSUMÉ

L'étalonnage est l'opération qui consiste à améliorer les estimations tirées d'une enquête infra-annuelle à l'aide des estimations correspondantes tirées d'une enquête annuelle. Par exemple, les estimations de l'enquête annuelle sur les ventes au détail peuvent servir à améliorer les estimations des ventes au détail mensuelles. Dans cet article, nous nous penchons tout d'abord sur le problème que pose l'étalonnage des séries chronologiques issues d'enquêtes économiques et nous analysons les solutions les plus appropriées dans les circonstances. Dans un deuxième temps, nous proposons deux nouvelles méthodes statistiques qui reposent sur un modèle non linéaire pour données infra-annuelles. Finalement, nous obtenons des estimations étalonnées en appliquant la méthode des moindres carrés pondérés et la méthode itérative du quotient de manière à assurer une certaine cohérence dans les tableaux de la série.

MOTS CLÉS: Erreurs d'enquête; modèle non linéaire; moindres carrés pondérés; ajustement proportionnel itératif.

1. INTRODUCTION

L'étalonnage a toujours été défini comme l'opération qui consiste à corriger des valeurs mensuelles ou trimestrielles tirées d'une source particulière en fonction de valeurs annuelles (données repères) tirées d'une autre source (voir Denton 1971, Cholette 1988a et Monsour et Trager 1979). Il peut s'agir, par exemple, de corriger les chiffres des expéditions mensuelles des manufacturiers canadiens de manière que leur somme égale la valeur des expéditions établie à l'aide de l'enquête annuelle sur les manufactures. L'étalonnage est aussi défini comme l'opération qui consiste à améliorer les estimations infra-annuelles tirées d'une source particulière à l'aide d'estimations annuelles tirées d'une autre source (voir Hillmer et Trabelsi, 1987). Contrairement à la première définition, la seconde suppose que les valeurs annuelles peuvent être erronées. Il peut s'agir, par exemple, d'améliorer les estimations des stocks mensuels des détaillants canadiens, établies à l'aide d'une enquête par sondage, au moyen des données sur les stocks de fin d'année tirées de l'enquête annuelle sur le commerce de détail. La seconde définition est celle qui s'applique le plus souvent aux séries économiques de Statistique Canada et c'est à elle que nous nous intéressons dans cet article.

Cet article se divise en quatre parties. Premièrement, nous allons exposer en détail le problème de l'étalonnage tel qu'il se présente pour la plupart des séries chronologiques de Statistique Canada issues des grandes enquêtes économiques. Ensuite, nous allons présenter et analyser les méthodes d'étalonnage les plus courantes qui peuvent s'appliquer à une série à la fois. Comme aucune de ces méthodes ne répond parfaitement aux exigences de Statistique Canada, nous proposons deux nouvelles méthodes statistiques qui peuvent s'appliquer à une série à la fois. Ces deux méthodes reposent sur un modèle non linéaire avec moindres carrés pondérés. Enfin, nous terminons cet exposé par l'étalonnage d'un tableau de séries chronologiques et l'étalonnage préliminaire.

2. POSITION DU PROBLÈME

Statistique Canada doit tenter d'améliorer des séries d'estimations infra-annuelles à l'aide de séries annuelles tirées des enquêtes-entreprises. Nous allons décrire ici les caractéristiques des données originales et les objectifs que nous poursuivons en ayant recours à l'étalonnage.

Les données infra-annuelles sont souvent entachées d'une erreur systématique à cause de problèmes de couverture. Citons tout d'abord le cas des nouvelles entreprises qui, souvent, sont intégrées à la base de sondage longtemps après être entrées en exploitation. C'est là que survient le sous-dénombrement. Citons aussi le cas des entreprises qui n'ont pas d'employés (il s'agit habituellement de petites entreprises) et qui, de ce fait, ne figurent pas dans la base infra- annuelle. Signalons enfin le problème du chevauchement des listes des grandes entreprises et des petites entreprises, qui servent de base de sondage pour les enquêtes infra-annuelles. Par conséquent, il y a de fortes chances que les estimations de totaux infra-annuels soient biaisées. Les données infra-annuelles ont aussi la particularité de provenir d'échantillons chevauchants. Il existe donc des covariances d'échantillonnage pour les estimations infra-annuelles qui se rapportent à des périodes différentes.

Pour ce qui a trait aux données annuelles, on peut supposer en pratique qu'elles sont sans biais puisqu'elles sont relativement peu touchées par le problème du chevauchement et que la base de sondage annuelle englobe les

[1] N. Laniel et K. Fyfe, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6

Figure 6: Diviseurs subjectifs utilisés pour le rajustement préalable des ventes d'autos à l'unité

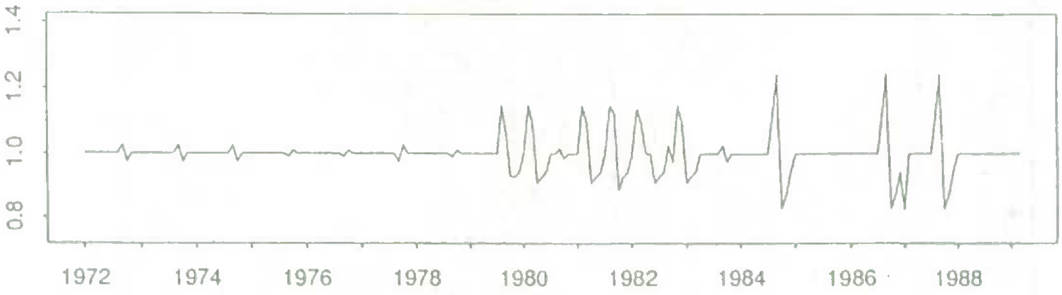


Figure 7: Diviseurs de la méthode REG-ARMMI utilisés pour le rajustement préalable des ventes d'autos à l'unité

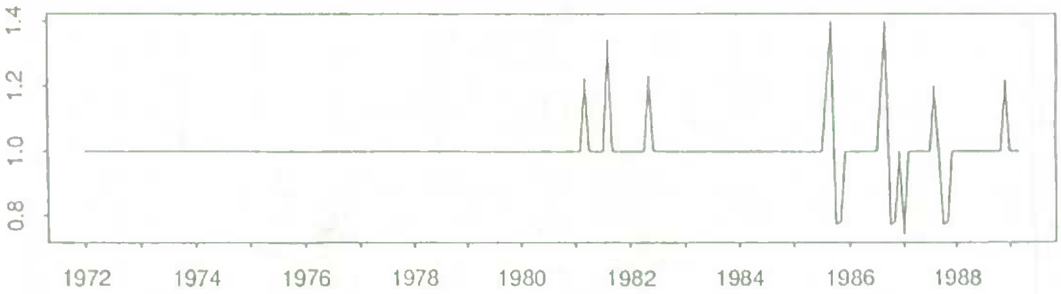


Figure 4: Facteurs liés aux jours ouvrables d'un trimestre (LJOT) pour les MAE

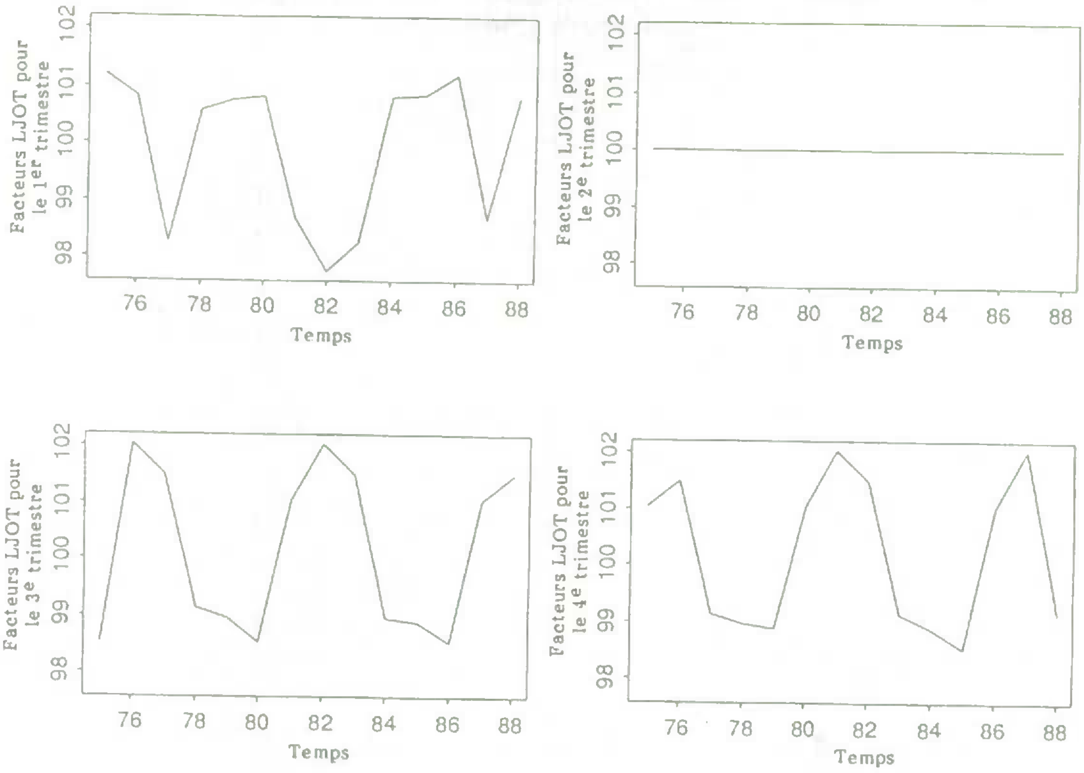


Figure 5: Ventes mensuelles d'automobiles à l'unité

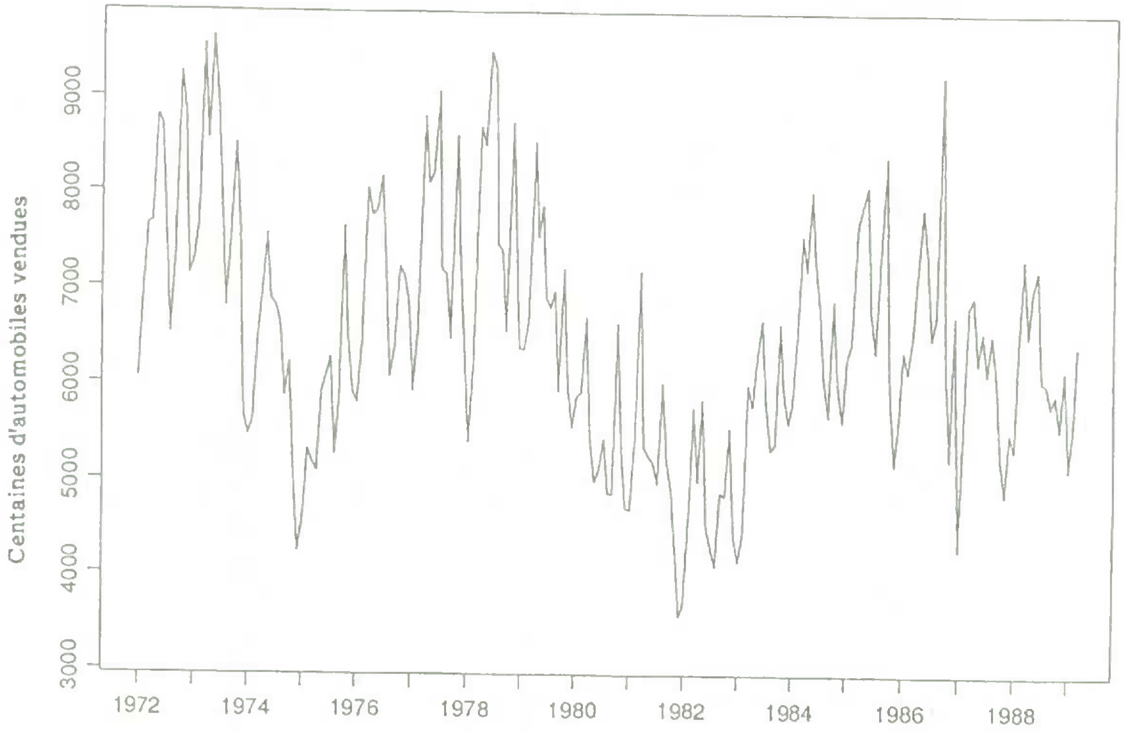


Figure 1: NVD

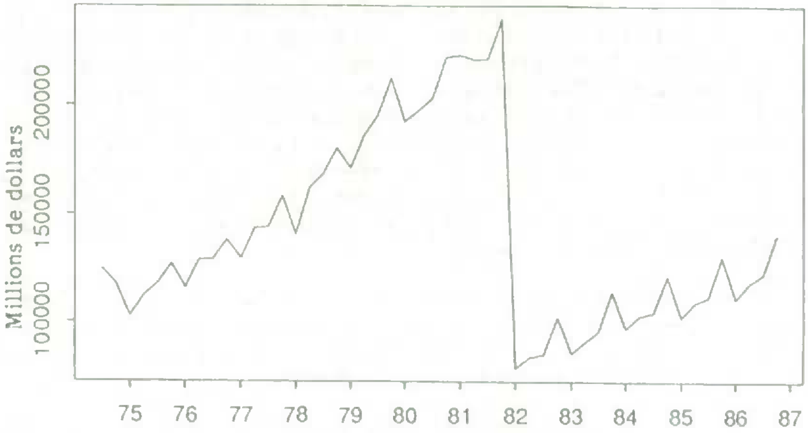


Figure 2: NCGAI

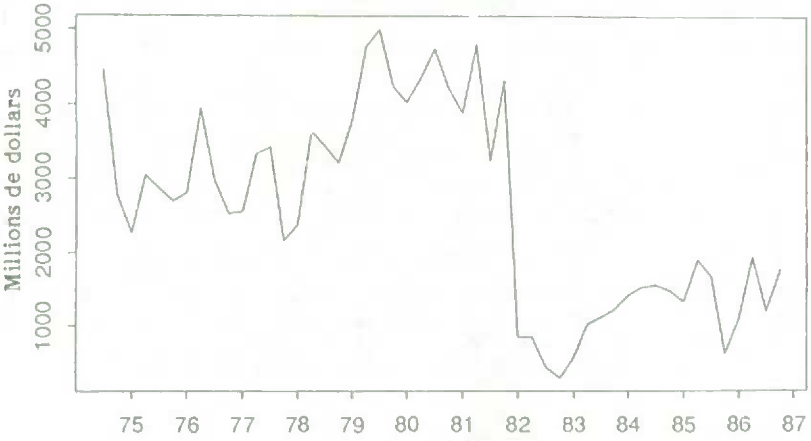
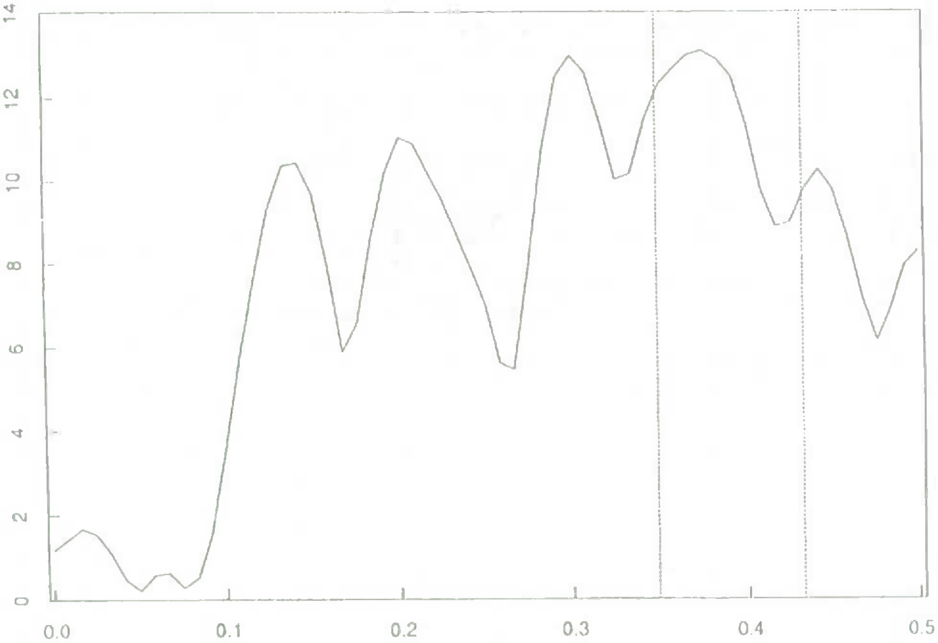


Figure 3: Périodogramme lissé de la série d'aléas provenant de l'IOCDE



Fréquence (cycles par mois)
Ligne pointillée = fréquence des jours ouvrables

(estimés par des spécialistes) aux diviseurs utilisés pour les rajustements préalables fondés sur un modèle. Les comparaisons de modèles REG-ARMMI nous offrent une façon objective de prendre de telles décisions, cas par cas, comme l'exemple ci-après le montre. La série de données sur les ventes d'automobiles à l'unité (VAU) de la figure 5 présente un certain nombre de mouvements extrêmes qui sont dûs à des campagnes de développement des ventes lancées par les fabricants d'automobiles. Ces campagnes ont servi à réduire les inventaires considérables des concessionnaires en offrant aux acheteurs des prêts à faible taux d'intérêt ou des remises en espèces. De telles campagnes de développement des ventes augmentent anormalement les ventes d'automobiles pour les mois au cours desquels elles sont en vigueur et causent une diminution inhabituelle du nombre de ventes au cours, approximativement, du mois qui suit la fin de la campagne. L'analyse de la série d'aléas tirés d'un rajustement des VAU par la méthode X-11-ARMMI a permis à un analyste d'obtenir les diviseurs utilisés pour les rajustements dont le graphique est donné à la figure 6.

Nous nous préoccupions du fait que la désaisonnalisation effectuée à l'aide de la méthode X-11-ARMMI et, par conséquent, la série d'aléas estimés qui en découlent, seraient compromises par les fluctuations qui résultent des campagnes de développement des ventes. Dans ce cas, les effets liés aux campagnes de promotion des ventes ne pouvaient être obtenus de façon fiable à partir des aléas.

Il nous a semblé plus opportun d'utiliser les procédures de détermination des valeurs aberrantes de la méthode X-12-ARMMI (voir Bell, 1983), avec certaines contraintes proposées par l'analyste, pour estimer les effets des campagnes de développement des ventes. Nous avons ajusté un tel modèle REG-ARMMI, avec des effets liés aux jours ouvrables, aux variables saisonnières fixes et à d'autres effets additifs des valeurs aberrantes, aux logarithmes de la série observée. Pour le modèle résultant (modèle 1), le graphique des effets estimés des valeurs aberrantes est donné à la figure 7. La valeur AIC pour ce modèle est $AIC_N^{(1)} = 3169.4$. La série de logarithmes rajustés en fonction de l'estimation de l'analyste (comme en 4. de la section 2) a aussi été ajustée à un modèle REG-ARMMI un peu différent (modèle 2) dont les variables de régression comprenaient le jour ouvrable, des variables saisonnières fixes et différentes valeurs aberrantes additives, qui, dans certains cas, allaient à l'encontre du rajustement effectué par l'analyste. Bien qu'on n'ait attribué aucun poids d'estimation de paramètre aux estimations de l'effet des campagnes de développement des ventes établies par l'analyste (nous ne savions pas comment procéder parce que les estimations n'ont pas été obtenues par estimation du maximum de vraisemblance), la valeur AIC pour ce modèle est beaucoup plus élevée, $AIC_N^{(2)} = 3187.0$. Nous concluons que le modèle 1 décrit mieux les données et, par conséquent, que les estimations des effets des campagnes de développement des ventes obtenues au moyen des termes de régression que renferme le modèle REG-ARMMI sont plus appropriées que celles provenant d'un examen de la série d'aléas qui découlent de l'emploi de la méthode X-11-ARMMI. D'autres analyses viennent aussi appuyer cette conclusion.

BIBLIOGRAPHIE

Bell, W. R. (1983). "A Computer Program for Detecting Outliers in Time Series." *Proc. Bus. Econ. Sec. ASA*, 634-639.

Bell, W. R. (1984). "Seasonal Decomposition of Deterministic Effects." Statistical Research Division Report No. CENSUS/SRD/RR-84/01. U.S. Bureau of the Census, Washington, D.C.

Bell, W. R. et Hillmer, S. C. (1983). "Modeling Time Series with Calendar Variation." *JASA* 78, 526-534.

Brockwell, P. J. et Davis, R. A. (1987). *Time Series: Theory and Methods.* New York: Springer Verlag.

Findley, D. F. (1988). "Comparing Not Necessarily Nested Models with the Minimum AIC and Maximum Kullback-Leibler Entropy Criteria: New Properties and Connections." *Proc. Bus. Econ. Sec. ASA*, 110-118.

Findley, D. F., Monsell, B. C., Otto, M. C. et Pugh, M. G. (1988). "Toward X-12-ARIMA." *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., pp. 591-624.

Findley, D. .F., Monsell, B. C., Shulman, H. .B et Pugh, M. G. (1990). "Sliding Spans Diagnostics for Seasonal and Related Adjustments." *JASA* 85 (à paraître).

identiques. L'emploi de variables saisonnières fixes est un mécanisme visant à nous permettre d'utiliser des comparaisons de modèles afin de décider si le comportement saisonnier de données plus récentes doit être estimé seulement à l'aide des données obtenues après 1981. Le tableau 1 donne les valeurs AIC pour les modèles REG-ARMMI avec ces deux types de variables explicatives ajustées aux séries du revenu net provenant des ventes au détail (NVD, voir la figure 1) et du revenu net du commerce de gros après impôts (NCGAI, voir la figure 2). Pour les deux comparaisons, la différence d dans le nombre de variables estimées est 3.

Tableau 1.: Essai des valeurs AIC pour un comportement saisonnier qui a changé.

	Mêmes variables saisonnières fixes (modèle 1)	Différentes variables saisonnières fixes (modèle 2)
NVD	1028.8	993.1
NCGAI	757.0	760.0

Ainsi, comme la figure 1 le laisse supposer, il y a un changement significatif dans la composante saisonnière du NVD en 1982, mais pas dans la composante saisonnière du NCGAI.

3.2 Test pour trouver la signification d'un effet indiqué.

Dans le cas des séries chronologiques mensuelles des importations aux Etats-Unis en provenance de la Communauté économique européenne, IOCDE, de janvier 1974 à décembre 1984 inclusivement, la valeur de la statistique F pour la régression portant sur les jours ouvrables, tirée du tableau de la méthode X-11-ARMMI avec (6, 124) degrés de liberté est 6.0. Cette valeur serait hautement significative si les hypothèses relatives à la régression menant à la distribution F étaient respectées. Cependant, la méthode X-11-ARMMI utilise une régression des moindres carrés ordinaires portant sur la série d'aléas estimés, qui est une série corrélée découlant de l'application d'une procédure de lissage, de sorte qu'une hypothèse fondamentale n'est pas respectée. Un modèle REG-ARMMI bien ajusté avec variables pour les jours ouvrables explique la corrélation. Nous ajustons trois modèles de ce genre à ces données, chacune avec une structure ARMMI $(0,1,1)(0,1,1)_{12}$, et avec les variables de régression suivantes:

a) terme constant, changement de niveau en février 1975 (modèle 1);
b) terme constant, changement de niveau en février 1975, variables pour les jours ouvrables (modèle 2);
c) terme constant, changement de niveau en février 1975, variables pour les jours ouvrables et variables pour le mois de février d'une année bissextile (modèle 3).

Les valeurs AIC correspondantes sont $AIC_N^{(1)} = 2241.9$, $AIC_N^{(2)} = 2250.1$ et $AIC_N^{(3)} = 2252.0$, de sorte que le modèle 1 est préféré, ce qui entraîne la production de la forme réduite de la statistique F de la méthode X-11-ARMMI. Un autre diagnostic, le périodogramme lissé de la série d'aléas donné dans la figure 3, n'a pas de sommmets aux fréquences des jours ouvrables, ce qui appuie la conclusion des analyses effectuées avec les modèles REG-ARMMI: la série n'a pas de composante des jours ouvrables significative.

3.3 Détection des effets liés aux jours ouvrables d'un trimestre.

On suppose depuis longtemps que, parce que le nombre de jours de la semaine au cours des trimestres varie beaucoup moins qu'au cours des mois civils, les effets liés aux jours ouvrables ne seraient pas significatifs pour des séries économiques trimestrielles. Cependant, Shelby Herman du bureau de l'analyse économique des Etats-Unis (U.S. Bureau of Economic Analysis) nous a récemment fait parvenir quelques séries de listes de paye qui, selon elle, avaient de tels effets. Nos analyses effectuées avec des modèles REG-ARMMI ont confirmé ses observations. Par exemple, nous avons ajusté trois modèles REG-ARMMI aux logarithmes de la série de listes de paye pour les MAE (fabricants de machines autres qu'électriques, du premier trimestre de 1975 au quatrième trimestre de 1988), avec des effets de régression qui comprenaient

a) aucun effet lié aux jours ouvrables d'un trimestre ou aux années bissextiles (modèle 1);
b) des effets liés aux jours ouvrables d'un trimestre (modèle 2), et
c) des effets liés aux jours ouvrables d'un trimestre et des effets dûs aux années bissextiles (modèle 3).

Les valeurs AIC pour les modèles correspondants sont $AIC_N^{(1)} = 808.8$, $AIC_N^{(2)} = 786.4$, et $AIC_N^{(3)} = 785.0$. Les modèles 2 et 3 sont tous deux préférés au modèle 1, et leurs effets estimés liés aux jours ouvrables sont presque identiques. Un graphique des facteurs liés aux jours ouvrables, qui sont les antilogarithmes des effets liés aux jours ouvrables du modèle 3 multipliés par 100, est donné à la figure 4.

4. COMPARAISON DES DIVISEURS SUBJECTIFS ET RELATIFS À LA MÉTHODE REG-ARMMI UTILISÉS POUR LES RAJUSTEMENTS PRÉALABLES

On nous a souvent demandé si l'on doit préférer les diviseurs subjectifs utilisés pour les rajustements préalables données mensuelles et m=4 pour les données trimestrielles.) Soit $I_t^{(j)}$ la variable indicatrice pour la j^e période, j=1, ..., m. (Par exemple, si m=4, alors $I_t^{(j)}=1$ si y_t représente la donnée statistique pour le j^e trimestre d'une année, et $I_t^{(j)}=0$ dans les autres cas.) Nous définissons alors

$$SF_t^{(j)} = I_t^{(m)} - I_t^{(4)},\ j=1,\ \ldots,\ m-1.$$

Le programme comprend aussi des variables de régression pour tenir compte de l'effet de Pâques sur les ventes au détail et de celui de plusieurs autres jours fériés aux E.-U.. Les utilisateurs peuvent introduire leurs propres variables de régression pour obtenir d'autres effets spéciaux ou pour tenir compte d'autres jours fériés liés au calendrier lunaire qui peuvent arriver pendant plusieurs mois de ce calendrier et avoir un effet économique, comme le Ramadan ou le nouvel an chinois.

Un ensemble spécial de variables explicatives que nous avons étudiées récemment ont été utilisées pour estimer les effets des jours ouvrables d'un trimestre.

Variables des jours ouvrables d'un trimestre et du premier trimestre d'une année bissextile

Soit $JT_t^{(j)}$ le nombre de jours de la semaine de type j (comme en 5. ci-dessus) dans le trimestre t. Nous définissons

$$JOT_t^{(j)} = JT_t^{(j)} - JT_t^{(7)},\ 1 \leq j \leq 6.$$

La variable explicative du premier trimestre d'une année bissextile $ABT1_t$ est définie en remplaçant les valeurs pour les mois de février dans la définition de FAB_t dans 6. ci-dessus par les valeurs pour les premiers trimestres.

3. EXEMPLES

Nous présentons maintenant certaines analyses basées sur les modèles REG-ARMMI qui utilisent les variables définies dans la section précédente.

La procédure du minimum de la valeur AIC de Akaike décrite plus haut sera utilisée quand il faudra comparer deux modèles entre lesquels on veut choisir. Quand le modèle 1 est une forme restreinte du modèle 2 avec moins de paramètres à estimer, cette procédure a une interprétation conventionnelle: on pourrait effectuer un test d'hypothèses avec comme hypothèse nulle que le modèle 1 est juste en supposant la distribution asymptotique chi carré du rapport de vraisemblance logarithmique,

$$H_0:\ 2(\hat{L}_N^{(2)} - \hat{L}_N^{(1)}) \sim X^2(d),$$

ce qui nous permet d'écrire $AIC_N^{(1)} - AIC_N^{(2)} \sim X^2(d) - 2d$. Il s'ensuit que la condition

$$AIC_N^{(1)} - AIC_N^{(2)} \geq 1$$

serait habituellement interprétée comme une différence statistiquement significative pour les valeurs AIC, ce qui favorise le modèle 2 (rejet de H_0).

3.1 Changement dans la définition des séries.

Dans le cadre d'un programme du gouvernement américain visant à diminuer le fardeau de déclaration des entreprises qui répondent aux enquêtes gouvernementales, une loi a été modifiée afin de réduire, à compter du premier trimestre de 1982, le nombre de sociétés devant répondre aux enquêtes réalisées pour produire le rapport financier trimestriel (Quarterly Financial Report). C'est pourquoi les niveaux de certaines des séries ont chuté brusquement d'une façon que les procédures d'estimation de la tendance qui font partie de la méthode X-11-ARMMI ne pouvaient traiter adéquatement; voir les figures 1 et 2 plus loin. Le fait que le segment d'après 1981 des séries pourrait avoir un comportement saisonnier différent de celui relevé pour le segment d'avant 1982, à cause du changement d'échantillon, constitue une préoccupation additionnelle. Pour étudier cette possibilité, nous avons ajusté à ces séries deux modèles REG-ARMMI entre lesquels on veut choisir. Les variables de régression de ces modèles comprenaient un changement de niveau au premier trimestre de 1982 et, soit un seul ensemble de variables saisonnières fixes pour les séries complètes (modèle 1), soit deux ensembles de telles variables (modèle 2), un ensemble pour le segment allant du premier trimestre de 1974 au quatrième trimestre de 1981 et l'autre pour le reste des séries. Cela signifie que, dans le cas du modèle 2, les coefficients pour l'effet saisonnier avant et après le changement de niveau peuvent être différents. Ainsi, si les retards AR et MM dans les modèles ajustés sont identiques, le modèle 1 constitue alors une forme restreinte du modèle 2 obtenue en exigeant que les deux ensembles de variables saisonnières fixes du modèle 2 aient des coefficients

$$\phi(B)\ (x_t - \beta z_t) = \theta(B)a_t \qquad (2.1)$$

où $\phi(B)$ et $\theta(B)$ sont des polynômes qui n'ont aucune racine dont la grandeur est inférieure à un, et a_t est un processus de bruit blanc qui n'est pas corrélé avec les valeurs précédentes de x_t. La méthode que nous utilisons pour estimer de tels modèles est décrite dans Findley et coll. (1988). Si $\hat{L}_N$ représente la valeur maximisée de la fonction de vraisemblance logarithmique tirée de N observations $x_1,\ldots x_N$ et si le nombre total de coefficients estimés dans $\phi(B)$, $\theta(B)$ et β est p, alors la statistique AIC de comparaison de Akaike pour le modèle ajusté est définie par

$$AIC_N = -2\hat{L}_N + 2p.$$

Quand on compare deux modèles estimés ou plus, on préfère habituellement le modèle avec la plus petite valeur AIC_N; voir, par exemple, Brockwell et Davis (1988) et Findley (1988). (La théorie n'appuie de telles comparaisons, faites à l'aide de la valeur AIC, que lorsque les polynômes $\phi(B)$ ont le même nombre de racines de grandeur 1 dans tous les modèles.)

La liste qui suit renferme sept ensembles typiques de variables de régression qui pourraient être incluses dans z_t et qui sont disponibles dans le module de prétraitement du programme X-12-ARMMI.

1. Valeur aberrante additive à t_0

$$VAA_t^{(t_0)} = \begin{matrix} 1, & t = t_0 \\ 0, & t \neq t_0 \end{matrix}$$

2. Changement de niveau à t_0

$$CN_t^{(t_0)} = \begin{matrix} 1, & t \geq t_0 \\ 0, & t < t_0 \end{matrix}$$

3. "Pente" entre t_0 et t_1

$$P_t^{(t_0,\ t_1)} = \begin{matrix} 1 & , t \geq t_1 \\ (t-t_0)/(t_1-t_0), & t_0 < t < t_1 \\ 0 & , t \leq t_0 \end{matrix}$$

4. Diviseur de rajustement préalable pour la série observée y_t

Supposons que $x_t = \log(y_t)$ et que D_t est un nombre positif par lequel on doit diviser y_t (par exemple, un déflateur ou une estimation, définie par l'utilisateur, de l'effet d'une campagne spéciale à court terme visant à développer les ventes), nous définissons

$$d_t = \log D_t$$

et nous donnons au coefficient de régression correspondant dans β la valeur 1, afin d'obtenir

$$x_t - d_t = \log(y_t/D_t).$$

5. Variables des jours ouvrables du mois

Si $JM_t^{(j)}$ représente le nombre de jours de la semaine de type j dans le mois t, avec $j=1,\ldots,7$ désignant lundi, ..., dimanche respectivement, nous définissons alors

$$JOM_t^{(j)} = JM_t^{(j)} - JM_t^{(7)} \quad , j=1,\ldots,6.$$

6. Variable pour le mois de février d'une année bissextile

$$FAB_t = \begin{matrix} -.25 \text{ pour le mois de février d'une année non bissextile} \\ .75 \text{ pour le mois de février d'une année bissextile} \\ 0 \text{ dans les autres cas} \end{matrix}$$

voir Bell et Hillmer (1983) et Bell (1984).

7. Variables saisonnières fixes

Soit m le nombre de périodes dans l'année au cours desquelles on obtient une observation. (Ainsi, m=12 pour les

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

PRÉTRAITEMENT À L'AIDE DE MODÈLES REG-ARMMI POUR LA DÉSAISONNALISATION

D.F. Findley et B.C. Monsell[1]

RÉSUMÉ

Le personnel qui étudie les séries chronologiques au sein de la division de la recherche statistique (Statistical Research Division) du bureau du recensement a élaboré des modules de logiciel qui peuvent être adaptés aux programmes de désaisonnalisation existants pour effectuer un pré- ou un post-traitement afin d'améliorer les fonctions de rajustement et de contrôle de la qualité. Le module de pré-traitement est un programme qui sert à la modélisation et à faire des estimations du maximum de vraisemblance "exactes" et efficientes, du point de vue calcul, de modèles ARMMI saisonniers, avec une fonction de régression par rapport à la moyenne. Le logiciel incorpore de nombreuses variables explicatives, afin de permettre à l'utilisateur de détecter et de modéliser diverses valeurs aberrantes et divers effets de calendrier courants que l'on retrouve dans les données économiques et que les programmes de désaisonnalisation existants ne peuvent traiter ou, ce qui est fréquent, traitent mal. Le programme permet aussi à l'utilisateur d'inclure ses propres variables explicatives. Le présent article renferme certains exemples qui illustrent l'emploi du module de prétraitement.

MOTS CLÉS: Modèle REG-ARMMI; AIC.

1. INTRODUCTION

Pour de nombreuses séries chronologiques économiques, l'établissement d'une procédure de désaisonnalisation appropriée exige plusieurs cycles de traitement afin d'effectuer des rajustements préalables et a posteriori. Le prétraitement comprend des prolongements des prévisions et des rajustements des données qui sont effectués, peut-être à titre d'essai, avant que les moyennes mobiles utilisées pour la désaisonnalisation réelle soient appliquées aux séries. On entend par post-traitement le calcul de divers diagnostics afin d'évaluer les effets, sur les séries désaisonnalisées, des options de prétraitement et de rajustement qui ont été choisies. L'objectif principal du post-traitement est de déterminer si l'on en est arrivé à un rajustement satisfaisant. Nous avons élaboré un nouvel ensemble de techniques de post-traitement, appelé analyse de périodes de temps mobiles (sliding spans analysis), qui est décrit dans Findley, Munsell, Shulman et Pugh (1990).

Le présent article porte sur le prétraitement. Nous présentons quatre exemples qui démontrent le rôle précieux de ce que nous appellerons les modèles REG-ARMMI (régression + ARMMI), pour déterminer ou comparer les rajustements préalables. Des fonctions permettant de déterminer et d'estimer les modèles REG-ARMMI tant typiques que personnalisés sont incluses dans le module de prétraitement d'un programme de désaisonnalisation, appelé provisoirement X-12-ARMMI, dont l'élaboration se termine au bureau du recensement américain; voir Findley, Munsell, Otto, Bell et Pugh (1988). Ce programme calcule aussi les diagnostics pour les périodes de temps mobiles.

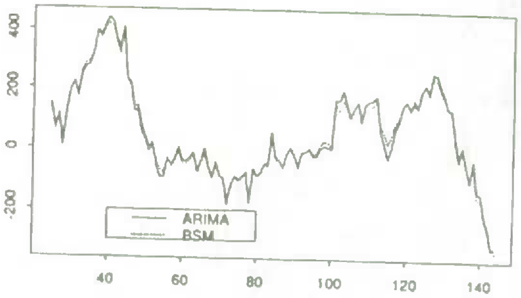
2. MODÈLES REG-ARMMI

De nombreuses séries chronologiques économiques montrent occasionnellement, au cours d'un bref intervalle, des mouvements erratiques considérables précédés et suivis de périodes plus longues de fluctuations raisonnablement stables. De telles perturbations peuvent être causées par des événements externes comme des grèves, des conditions climatiques extrêmes, des hostilités internationales et des changements dans les politiques gouvernementales, ou elles peuvent découler de facteurs internes comme des changements dans la classification économique ou dans l'échantillon utilisé pour définir ou pour obtenir les séries. Les perturbations de ce genre, particulièrement celles qui entraînent un changement durable dans le niveau des séries, compromettent la fiabilité des désaisonnalisations obtenues à l'aide de la méthode X-11-ARMMI et des procédures connexes, et elles rendent plus difficile la détermination des modèles ARMMI utilisés pour prévoir de telles séries.

Il est souvent possible de modéliser ces perturbations de façon adéquate au moyen de modèles REG-ARMMI,ce que nous décrirons maintenant. Soit x_t la série à modéliser (c'est souvent le logarithme de la série observée y_t), B l'opérateur de retard, $Bx_t = x_{t-1}$, et z_t un vecteur de variables de régression connues dont le vecteur des coefficients β peut renfermer des coefficients connus et des coefficients inconnus. Les coefficients inconnus seront calculés sous forme d'un sous-vecteur d'estimations gaussiennes du maximum de vraisemblance des paramètres inconnus d'un modèle REG-ARMMI, ce qui correspond à un modèle de séries chronologiques de la forme

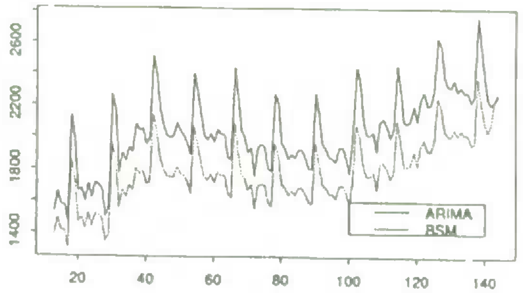
[1] Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

Chômage chez les adolescents



Estimations d'extraction des signaux

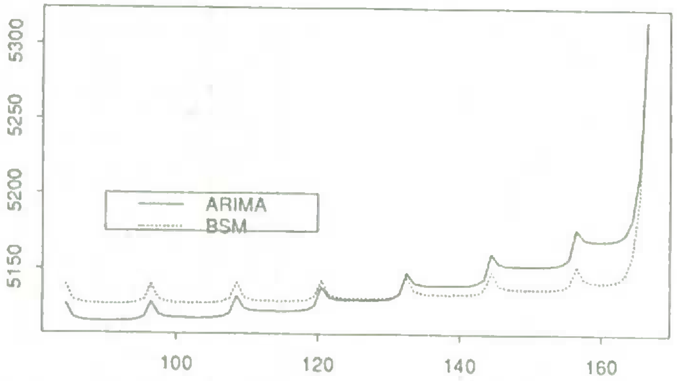
Figure 3.a



Variances d'extraction des signaux

Figure 3.b

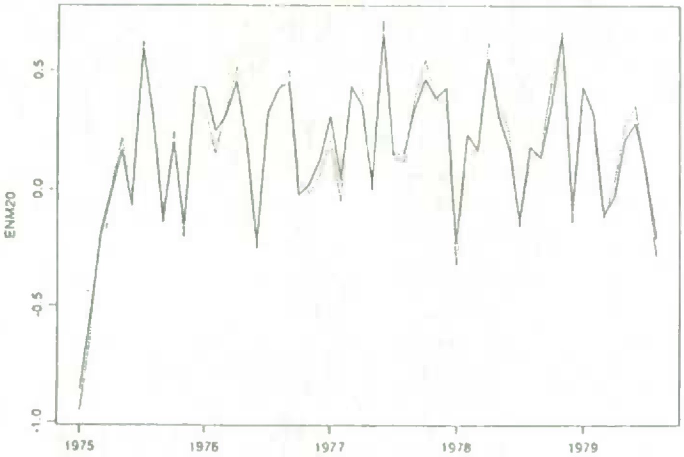
Total des démarrages de construction de 5 maisons ou plus aux États-Unis



Variances d'extraction des signaux

Figure 4

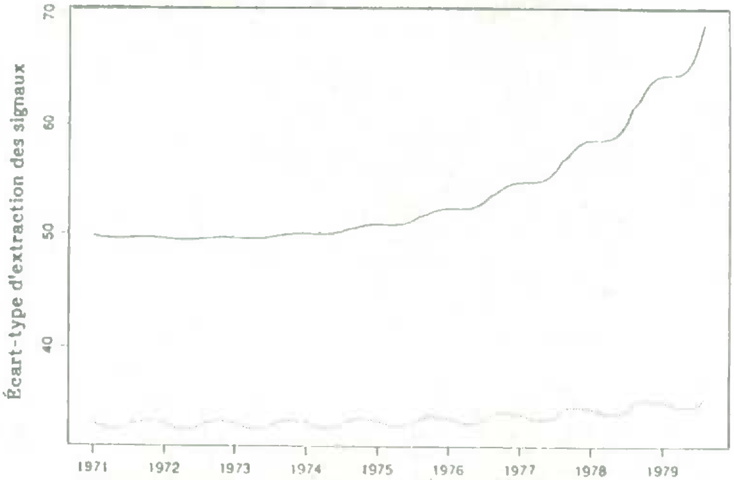
Pourcentage de variation mensuelle, données dont les variations saisonnières ont été corrigées (après 1975)



Ligne pleine = Correction des variations saisonnières canoniques
Ligne pointillée = Correction des variations saisonnières BSM

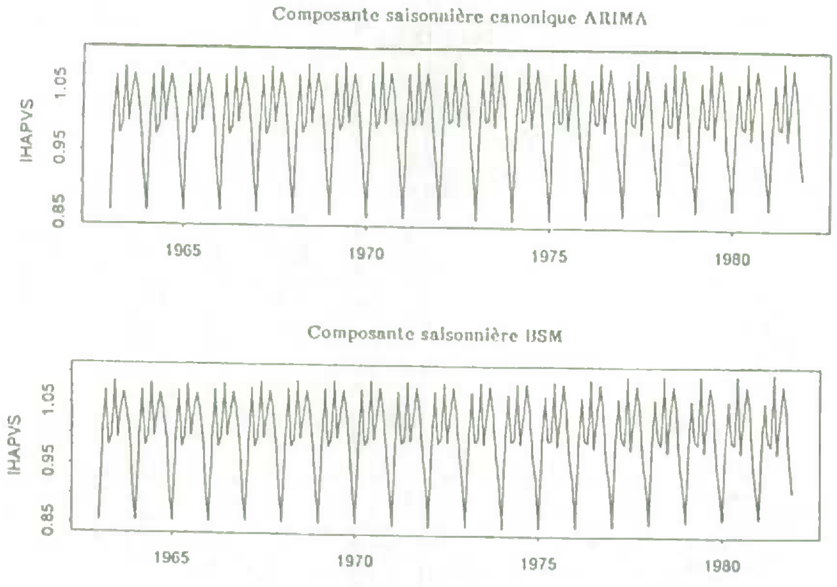
Figure 2.b

ENM20, écart-type d'extraction des signaux de données dont les variations saisonnières ont été corrigées (après 1971)



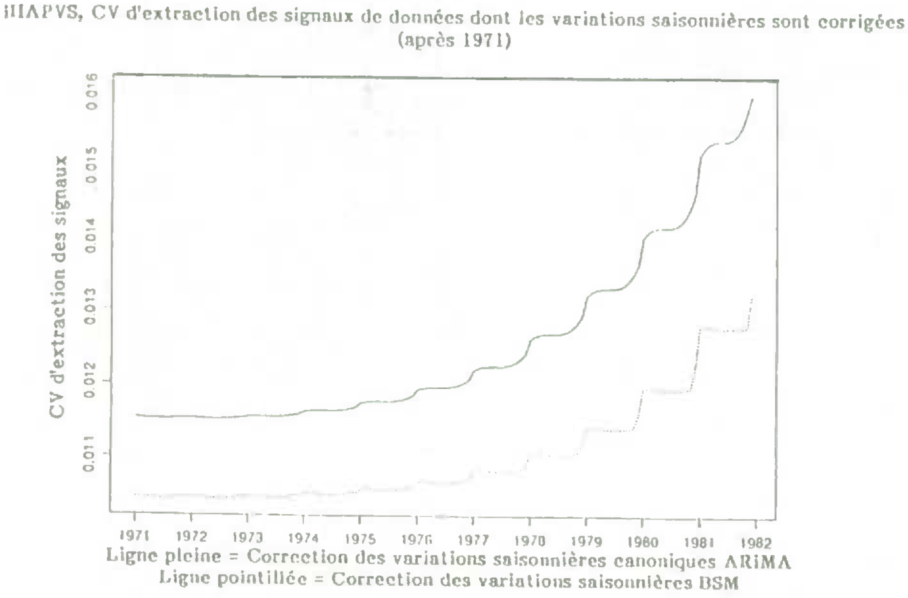
Ligne pleine = Correction des variations saisonnières canoniques ARIMA
Ligne pointillée = Correction des variations saisonnières BSM

Figure 2.c


Composante saisonnière canonique ARIMA

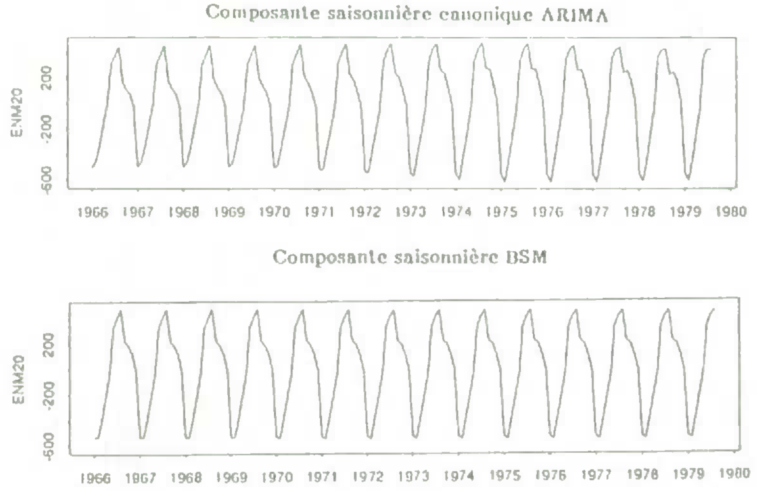
Composante saisonnière BSM

Figure 1.a


IHAPVS, CV d'extraction des signaux de données dont les variations saisonnières sont corrigées (après 1971)

Ligne pleine = Correction des variations saisonnières canoniques ARIMA
Ligne pointillée = Correction des variations saisonnières BSM

Figure 1.b


Composante saisonnière canonique ARIMA

Composante saisonnière BSM

Figure 2.a

_____(1987a), "Initializing the Kalman Filter for Nonstationary Time Series Models", Research Report Number 87/33, Statistical Research Division, Bureau of the Census.

_____ (1987b), "Time Series Methods for Survey Estimation", Research Report Number 87/20, Statistical Research Division, Bureau of the Census.

_____(1989), "Modeling Time Series Subject to Sampling Error," Research Report Number 89/01, Statistical Research Division, Bureau of the Census.

Bell, W.R. et Pugh, M.G. (1990) "Alternative Approaches to the Analysis of Time Series Components", Research Report Number 90/01, Statistical Research Division, Bureau of the Census.

Binder, D. A. et Dick, J. P. (1989) "Modelling and Estimation for Repeated Surveys", *Survey Methodology*, 14, à publier.

Box, G.E.P. et Jenkins, G. M. (1976), *"Time Series Analysis : Forecasting and Control"*, San Francisco: Holden Day.

Burman, J. P. (1980), "Seasonal Adjustment by Signal Extraction", *Journal of the Royal Statistical Society Series A*, 143, 321- 337.

Burman, J. P. et Otto, M. (1988), "Outliers in Time Series", Research Report Number 88/14, Statistical Research Division, Bureau of the Census.

Buys Ballot, C. H. D. (1847) *Les Changements Périodiques de Température*, Utrecht : Kemink et Fils.

Carlin, J. B. et Dempster, A. P. (1989) "Sensitivity Analysis of Seasonal Adjustments : Empirical Case Studies", *Journal of the American Statistical Association*, 84, 6-20.

Findley, D. F. (1983), "Comments on 'Comparative Study of the X-11 and BAYSEA Procedures of Seasonal Adjustment' by H. Akaike and M. Ishiguro", *Applied Time Series Analysis of Economic Data*, ed. Arnold Zellner, Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

_____(1988), "Comparing Not Necessarily Nested Models With the Minimum AIC and the Maximum Kullback-Leibler Entropy Criteria: New Properties and Connections", Research Report Number 88/21, Statistical Research Division, Bureau of the Census.

Findley, D. F., Monsell, B. M., Otto, M. C. Bell, W. R. et Pugh M. G. (1988), "Toward X-12 ARIMA", Proceedings of the Fourth Annual Research Conference, U. S. Department of Commerce, Bureau of the Census.

Gersch, W. et Kitagawa, G. (1983), "The Prediction of Time Series With Trends and Seasonalities", *Journal of Business and Economic Statistics*, 1, 253-264.

Harvey, A. C. (1985), "Trends and Cycles in Marcroeconomic Time Series", *Journal of Business and Economic Statistics*, 3, 216-227.

Hillmer, S. C., Bell, W. R. et Tiao, G. C. (1983a), "Modeling Considerations in the Seasonal Adjustment of Economic Time Series", *Applied Time Series Analysis of Economic Data*, ed. Arnold Zellner, U.S. Department of Commerce, Bureau of the Census, 74-100.

Hillmer, S. C. et Tiao, G. C. (1982), "An ARIMA-Model-Based Approach to Seasonal Adjustment", *Journal of the American Statistical Association*, 77, 63-70.

Hotta, L. K. (1989), "Identification of Unobserved Components Models", *Journal of Time Series Analysis*, 10, 259-270.

Kitagawa, G. et Gersch, W. (1984), "A Smoothness Priors-State Space Modeling of Time Series With Trend and Seasonality", *Journal of the American Statistical Association*, 79, 378-389.

Kohn, R. et Ansley, C. F. (1986), "Estimation, Prediction, and Interpolation for ARIMA Models With Missing Data", *Journal of the American Statistical Association*, 81, 751- 761.

Maravall, A. (1985), "On Structural Time Series Models and the Characterisation of Components", *Journal of Business and Economic Statistics*, 3, 350-355.

Nerlove, M., Grether, D. M. et Caravallo, J. L. (1979), *Analysis of Economic Time Series: A Synthesis*, New York: Academic Press.

Prothero, D. L. et Wallis, K. F. (1976) "Modeling Macroeconomic Time Series", *Journal of the Royal Statistical Society Series A*, 139, 468-500.

Pfeffermann, D. (1989) "Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys", paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Scott, A. J. et Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods", *Journal of the American Statistical Association*, 69, 674-678.

Scott, A. J., Smith, T.M.F. et Jones, R. G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys", International Statistical Review, 45, 13-28.

Watson, M. W. (1987) "Uncertainty in Model-Based Seasonal Adjustment Procedures and Construction of Minimax Filters", Journal of the American Statistical Association, 82, 395-408.

$$(1-B)(1-B^{12})s_t = (1 - .47B)(1 - .89B^{12})a_t \text{ , } \sigma_a^2 = .0215 .$$

Nous avons utilisé les modèles ci-dessus dans l'estimation de l'extraction des signaux de s_t, puis nous avons fait la même chose avec un BSM ajusté pour s_t, cela avec les mêmes modèles e_t que ceux donnés ci-dessus. Le modèle BSM s'est relativement bien ajusté à ces deux séries car DAIC = AIC(BSM) -AIC(ARIMA) = -3,1 pour le chômage chez les adolescents et DAIC = 1,8 pour les démarrages de construction de maisons. (La véracité de ces AIC peut être discutée car les modèles e_t ne sont pas ajustés avec les données des séries chronologiques.) La figure 3.a. indique les estimations des points d'extraction des signaux pour le chômage chez les adolescents dans le cas des deux modèles; $(1-B^{12})\hat{s}_t$ est inclut afin d'éliminer les effets obscurcissants de la saisonnalité. Le BSM donne une estimation de la variance du signal inférieure à celle du modèle ARIMA et, pour cette raison, fournit des estimations légèrement plus régulières. La figure 3.b. illustre des différences importantes dans les variances d'extraction des signaux des deux modèles. Les deux estimations d'extraction des signaux de la série des démarrages de construction de maisons étaient pratiquement identiques et ne sont pas indiquées. La figure 4 donne les coefficients de variation de l'extraction des signaux (écart-type pour le logarithme de la série) pour la dernière moitié de la série des démarrages de construction de maisons - ceux de la première moitié seraient identiques. Bien qu'il y ait quelques différences intéressantes dans les caractéristiques, l'amplitude des différences est petite.

5. CONCLUSIONS

Même les conclusions tirées à la section 2 doivent être quelque peu provisoires. Il serait intéressant que des études semblables soient entreprises avec d'autres ensembles de séries chronologiques. En raison du nombre limité d'exemples étudiés aux sections 3 et 4, les conclusions les concernant ne peuvent être que des suggestions. En résumé :

1. Les données permettent fréquemment de distinguer les modèles ARIMA des modèles de composantes. Dans le cas des 45 séries analysées, l'AIC a fortement favorisé les modèles ARIMA par rapport aux BSM. Dans la mesure où l'ajustement d'un modèle est important, supposer uniquement que le BSM fournit un ajustement adéquat serait dangereux.

2. Nous avons constaté qu'il était plus difficile d'ajuster les modèles de composantes que les modèles ARIMA. Bien que nous aurions aimé constater que l'ajout d'une composante AR stationnaire ou d'un autre terme cyclique améliore les ajustements des modèles de composantes, nous avons été incapables d'ajuster ces modèles par suite de difficultés d'ordre numérique.

3. Les estimations de points d'extraction des signaux pour les corrections des variations saisonnières et les estimations des enquêtes au moyen des modèles ARIMA et des BSM ont été très peu différentes dans les exemples étudiés. Les variances d'extraction de signaux varient beaucoup plus, bien que dans les exemples de correction des variations saisonnières, les variances correspondant aux deux modèles puissent être considérées assez petites. Cette dernière question devrait être étudiée de façon plus approfondie afin que l'on puisse déterminer si les variances des corrections des variations saisonnières des modèles avec décompositions canoniques ou approximativement canoniques sont normalement très petites.

BIBLIOGRAPHIE

Abraham, B. et Box, G.E.P. (1978), "Deterministic and Forecast- Adaptive Time-Dependent Models", *Applied Statistics*, 27, 120-130.

Akaike, H. (1973) "Information Theory and an Extension of the Likehood Principle", *2nd International Symposium on Information Theory*, ed. B. N. Petrov et F. Czaki, Budapest: Akademia Kiado, 267-287.

_____ (1980), "Seasonal Adjustment by a Bayesian Modeling", *Journal of Time Series Analysis*, 1, 1-13.

Bell, W. R. (1984) "Seasonal Decomposition of Deterministic Effects", Research Report Number 84/01, Statistical Research Division, Bureau of the Census.

Bell, W.R. (1987) "A Note on Overdifferencing and the Equivalence of Seasonal Time Series Models With Monthly Means and Models With $(0,1,1)_{12}$ Seasonal Parts When = 1", *Journal of Busisness and Economic Statistics*", 5, 383-387.

Bell, W.R. et Hillmer, S.C. (1983), "Modeling Time Series with Calendar Variation", *Journal of the American Statistical Association*, 78, 526-534.

_____ (1984), "Issues Involved with the Seasonal Adjustment of Economic Time Series", (avec discussion), *Journal of Business and Economic Statistics*, 2, 291-320.

La figure 2.a donne les composantes saisonnières ARIMA et BSM pour ENM20. Nous pouvons cette fois-ci voir une différence : la composante saisonnière ARIMA évolue régulièrement en fonction du temps, tandis que la composante BSM reste relativement fixe. (Dans le cas du BSM de ENM20, $\hat{\sigma}_1^2 = 27$ et $\hat{\sigma}_2^2 = 16\ 500$.) La figure 2.b illustre les résultats de la correction des variations saisonnières pendant les cinq dernières années au cours desquelles les données ont été observées. Bien que l'on puisse constater des différences, celles-ci ne peuvent pas être importantes car les variations mensuelles ne sont pas suffisamment grandes et dépassent rarement 0,5 %. La figure 2.c indique des variations encore plus grandes que dans IHAPVS en ce qui concerne les écarts-types d'extraction des signaux. Les écarts-types du BSM augmentent très peu à la fin de la série puisque une composante saisonnière essentiellement fixe est estimée. Enfin, la caractéristique la plus intéressante de la figure 2.c est peut-être le fait que les écarts-types sont très petits par rapport aux valeurs observées des séries qui vont de 40 000 à 50 000.

Nous avons supposé que $\mathrm{Var}(S_t - \hat{S}_t) \to 0$ lorsque $\theta_{12} \to 1$ dans le modèle ARIMA et lorsque $\sigma_1^2 \to 0$ dans le BSM, ce qui explique probablement les petits écarts-types d'extraction des signaux observés dans les deux exemples. Toutefois, si nous posons $\theta_{12} = 1$ ou $\sigma_1^2 = 0$ et que nous utilisions plutôt un modèle dans lequel les effets de la régression saisonnière sont fixes, les variances d'extraction des signaux ne seront pas nulles étant donné que nous aurons introduit une erreur dans l'estimation des paramètres de la régression saisonnière. Un aspect curieux de ces résultats est la discontinuité apparente entre les résultats obtenus avec $\theta_1 < 1$ (ou $\sigma_1^2 > 0$) et $\theta_{12} = 1$ (ou $\sigma_1^2 = 0$).

4. ESTIMATION D'ENQUÊTES RÉPÉTÉES

Scott et Smith (1974) et Scott, Smith et Jones (1977) ont proposé d'utiliser des techniques d'extraction des signaux des séries chronologiques pour estimer des enquêtes périodiques. Soient s_t, l'effectif réel de la population (le signal) et e_t, l'erreur d'échantillonnage au temps t. Nous utilisons l'extraction des signaux pour estimer s_t dans :

$$Y_t = S_e + e_t . \tag{4.1}$$

Si Y_t est le logarithme de la série originale, $\exp(s_t)$ et $\exp(e_t)$ sont l'effectif réel de la population et l'erreur d'échantillonnage multiplicatif dans la série originale. N'importe lequel des modèles traités à la section 2 peut être utilisé pour s_t. Par exemple, Binder et Dick (1989) et Bell et Hillmer (1989) utilisent des modèles ARIMA, tandis que Pfefferman (1989) utilise un BSM. Normalement, tous les termes de régression du modèle font aussi partie de s_t.

La création de modèles en vue d'estimer une enquête est traitée dans les références citées ci-dessus. Une première distinction entre cette application et ce que nous avons étudié auparavant est que le modèle de e_t est généralement estimé d'une façon ou d'une autre à l'aide de microdonnées d'enquête. Le modèle de l'erreur d'échantillonnage est maintenu fixe lorsque l'on estime les paramètres du modèle s_t en utilisant des données de séries chronologiques sur Y_t. On est alors amené à se poser des questions au sujet de la sensibilité des résultats d'estimation d'enquête à n'importe lesquels des aspects de la modélisation. Nous allons examiner la sensibilité des résultats au choix entre un modèle ARIMA et un BSM pour s_t.

Soient deux séries chronologiques. Pour la première, le chômage chez les adolescents américains (en milliers d'unités) entre janvier 72 et décembre 83, Bell et Hillmer (1987b) ont développé le modèle suivant pour $Y_t = s_t + e_t$:

$$(1-B)(1-B^{12})s_t = (1 - .27B)(1 - .68B^{12})a_t\ , \quad \sigma_a^2 = 4294,$$

$$e_t = h_t\tilde{e}_t \quad (1 - .6B)\tilde{e}_t = (1-.3B)c_t\ , \quad \sigma_c^2 = .8767\ , \quad h_t^2 = -.0000153\ Y_t^2 + 1.971\ Y_t$$

Le modèle de s_t a été réestimé, ce qui a permis d'obtenir des valeurs de paramètres légèrement différentes de celles indiquées dans Bell et Hillmer (1987b). Avec $\sigma_c^2 = 0{,}8767$ et $\mathrm{Var}(\tilde{e}_t) = 1$, h_t est l'écart-type (estimé) des erreurs d'échantillonnage et il varie avec le temps. La modélisation de la deuxième série, au moins 5 démarrages de construction de maisons américaines, est très semblable à celle de la série des démarrages de construction de maisons unifamiliales américaines également traitée dans Bell et Hillmer (1987b). Les erreurs d'échantillonnage de cette série semblent être approximativement indépendantes du temps car elles ont une variance relative égale à 0,00729, ce qui est aussi la variance approximative du logarithme des erreurs d'échantillonnages multiplicatives. Le modèle ARIMA estimé pour le signal du logarithme de la série chronologique est :

base en composantes. Nous pouvons tenir compte de cette incertitude dans la décomposition des modèles de composantes en définissant une décomposition canonique analogue à celle des modèles ARIMA, c'est-à-dire en soustrayant le plus possible de bruit blanc de S_t et en l'ajoutant à N_t par l'intermédiaire de I_t. Dans Bell et Pugh (1990, annexe A.1), nous avons montré que la décomposition canonique résultante des modèles de composantes $Z_t = \bar{S}_t + \bar{N}_t = \bar{S}_t + [T_t + \bar{I}_t]$, a une composante irrégulière canonique $\bar{I}_t$ dont la variance est $\bar{\sigma}_3^2 = \sigma_3^2 + \sigma_1^2/144$ et une composante saisonnière canonique $\bar{S}_t$ qui est conforme au modèle :

$$U(B)\bar{S}_t = \psi(B)\bar{\varepsilon}_{1t} \text{ , } \bar{\varepsilon}_{1t} \sim \text{iid } N(0,\bar{\sigma}_1^2) \qquad (3.1)$$

où $\psi(B)$, d'ordre 11, est donné dans le tableau 2 et $\bar{\sigma}_1^2 = .8081\ \sigma_1^2$. (Bell et Pugh (1990) ont aussi traité de la tendance canonique pour les modèles de composantes.) En fait, il s'agit de la même forme que le modèle saisonnier canonique de Burman (1980) et d'Hillmer et Tiao (1982), bien que leur modèle saisonnier ait généralement un $\psi(B)$ et un $\bar{\sigma}_1^2$ différents (ceci dépend du modèle ARIMA). Comme avec les modèles ARIMA, l'emploi de n'importe quelle autre décomposition admissible (correspondant à n'importe quelle décomposition valide de la fonction génératrice de covariance), y compris celle définie par le modèle de composantes initial ajusté, peut être considéré comme un ajout de bruit blanc à la composante saisonnière canonique $\bar{S}_t$. On remarquera que pour un modèle de composantes donné, le modèle de $\bar{S}_t$ dans (3.1) peut être obtenu de façon triviale. En outre, l'extraction des signaux de la désaisonnalisation canonique peut s'effectuer de la façon habituelle avec un ajusteur de Kalman utilisant le modèle (3.1) pour $\bar{S}_t$ et faisant augmenter la variance irrégulière à $\bar{\sigma}_3^2$.

Table 2 : Coefficients ψ_k pour $\psi_1 B - \cdots - \psi_{11} B^{11}$

k	ψ_k	k	ψ_k	k	ψ_k
1	0,205555	5	0,100648	9	0,031188
2	0,175919	6	0,080059	10	0,018953
3	0,148557	7	0,061661	11	0,008593
4	0,123471	8	0,045395		

On remarquera que la quantité de variance éliminée de la composante saisonnière du modèle de composantes, soit $\sigma_1^2/144$, sera petite à moins que σ_1^2 soit grand par rapport à σ_2^2 et σ_3^2. Toutefois, l'opposé est également vrai pour la série traitée ici : $\sigma_1^2/(\sigma_2^2 + \sigma_3^2)$ est supérieur à 0,07 pour seulement deux des 45 séries. Ceci a deux conséquences : (1) le modèle de composantes estimé suppose habituellement une saisonnalité pratiquement fixe, et (2) la décomposition du modèle de composantes original sera souvent très proche de la décomposition du modèle de composantes canonique. En fait, dans les séries que nous allons étudier, les corrections des variations saisonnières des décompositions du modèle de composantes canoniques et du modèle de composantes original ont été pratiquement identiques. Comme le choix de la décomposition semble avoir peu d'effet, nous n'étudierons pas cet aspect de façon plus approfondie. Cela ne signifie pas non plus que si l'on choisit une décomposition autre que la décomposition canonique, il n'y aura pas d'effets importants, mais nous n'allons pas étudier cette question dans le présent article.

Pour examiner les différences possibles des corrections de la variation saisonnière en fonction du choix du modèle, nous avons examiné de telles corrections pour deux séries : IHAPVS (valeur des appareils ménagers américains livrés entre janvier 1962 et décembre 1981) et ENM20 (milliers d'employés mâles de 20 ans et plus dans les secteurs autres que l'agriculture entre janvier 65 et août 79), une série analysée par Bell et Hillmer (1984). IHAPVS est l'une des séries pour lesquelles le BSM s'ajuste le mieux (DIAIC = -7), tandis que l'ajustement du BSM pour ENM20 était plutôt médiocre (DAIC = 13,7), sans être toutefois le pire. Comme les logarithmes n'ont pas été utilisés dans ENM20, il a été possible de faire une décomposition additive.

La figure 1.a donne les composantes saisonnières estimées de l'ARIMA et du BSM pour IHAPVS. Il est nécessaire de faire une analyse approfondie pour détecter les différences. Comme il en va de même des corrections des variations saisonnières, nous ne les avons pas présentées. La figure 1.b donne les écarts-types des extractions de signaux pour IHAPVS, exprimés sous forme de coefficients de variation. On constate des différences importantes car les CV du modèle ARIMA augmentent d'au moins 20 % vers la fin de la série. (Il est bon de noter que les résultats du modèle ARIMA ne sont pas nécessairement mauvais.) Toutefois, les CV peuvent tous être considérés comme petits car aucun ne dépasse 1,6 %.

Nous avons constaté que les modèles de composantes étaient bien plus difficiles à ajuster que les modèles ARIMA réguliers. Par exemple, l'obtention de bonnes valeurs de départ pour l'itération non linéaire à appliquer aux paramètres des modèles de composantes semble importante, ce qui n'est pas le cas avec les paramètres des modèles ARIMA. Nous n'avons présenté aucun résultat pour les modèles ayant une quatrième composante, modèles semblables à celui de (2.6), car nous avons été incapables d'ajuster convenablement de tels modèles. L'ajout d'une quatrième composante a fait sortir la recherche non linéaire à l'extérieur de la région de stationnarité de V_t, ce qui a provoqué l'arrêt du programme avec toutes les séries. Bien qu'il soit possible de contourner ce problème par programmation et bien que l'inclusion d'une quatrième composante puisse améliorer les ajustements, nous avons estimé que les difficultés rencontrées étaient décourageantes. Sans avoir entrepris une étude formelle des problèmes numériques que nous avons rencontrés avec les modèles de composantes, ceux-ci semblaient dus au fait que la vraisemblance était plutôt plate dans certaines directions de l'espace des paramètres. Pour cette raison, nous estimons que les avantages de simplicité et d'interprétabilité des modèles de composantes fréquemment mis de l'avant ne semblent pas correspondre à la réalité.

Les difficultés de calcul que nous avons constatées semblent suggérer une explication finale possible de nos résultats, à savoir que notre logiciel ne fonctionne pas convenablement et qu'il ne maximise pas véritablement la vraisemblance. Bien que nous ayons soigneusement vérifié notre programme, il nous est impossible d'éliminer cette possibilité en toute certitude. Nous serons très heureux de fournir nos données à toute personne intéressée à vérifier nos résultats. Nous serions même encore plus intéressés à ce qu'une étude portant sur d'autres séries soit entreprise afin de déterminer si elle obtient ou non des résultats semblables aux nôtres.

3. DÉSAISONNALISATION

Bien que la section 2 semble indiquer que les modèles ARIMA puissent bien mieux s'ajuster à une série chronologique que les modèles de composantes, il reste encore à déterminer la différence pratique résultant du choix d'un modèle plutôt qu'un autre. Nous avons étudié l'effet du choix d'un modèle sur la désaisonnalisation. Avec un modèle de composantes donné, on peut effectuer la correction saisonnière en appliquant un ajusteur Kalman à la série (voir p. ex. Gersch et Kitagawa 1983). Avec les modèles ARIMA, il est nécessaire de poser suffisamment d'hypothèses pour passer de modèles ARIMA simples portant sur des séries observées à des modèles de composantes uniques. Cette question a été traitée par Burman (1980) et par Hillmer et Tiao (1982), qui ont étudié une gamme de décompositions possibles et ont proposé un choix conduisant à une décomposition unique en des modèles de composantes. (Les deux approches diffèrent quelque peu pour certains modèles peu fréquents.) Les hypothèses sous-jacentes sont posées et traitées de façon plus approfondie par Bell et Hillmer (1984). Nous verrons un peu plus loin que nous pouvons également envisager une gamme de décompositions pour n'importe quel modèle de composantes.

Dans le cas où Y_t est conforme à (2.1) et (2.3), Burman (1980) et Hillmer et Tiao (1982) ont effectué une décomposition de (2.2) en décomposant en fractions partielles la fonction génératrice de covariance (CGF) de Z_t, $\gamma_Z(B)$, ce qui leur a permis d'obtenir les $\gamma_S(B)$, $\gamma_T(B)$, $\gamma_I(B)$ et des CGF, ainsi que les modèles ARIMA des composantes. On obtient ainsi une gamme de décompositions admissibles correspondant à $\gamma_Z(B) = [\gamma_S(B) - \gamma_1] + [\gamma_T(B) - \gamma_2] + [\gamma_I(B) + \gamma_1 + \gamma_2]$, pour toutes valeurs γ_1 et γ_2, telle que chaque terme entre crochets est ≥ 0 pour tout $B = e^{i\lambda}$. Cette gamme reflète l'incertitude sous-jacente de la décomposition; on obtient une décomposition donnée pouvant être utilisée pour la désaisonnalisation en spécifiant γ_1 et γ_2. Burman (1980) et Hillmer et Tiao (1982) proposent de choisir les γ_1 et γ_2 ($\bar{\gamma}_1 = \min_\lambda \gamma_S(e^{i\lambda})$ et $\bar{\gamma}_2 = \min_\lambda \gamma_T(e^{i\lambda})$) maximaux admissibles, de façon à obtenir une décomposition dite canoniques qui possède plusieurs propriétés intéressantes. Si l'on s'intéresse en particulier à la décomposition saisonnière-non saisonnière, les composantes correspondant à n'importe quel γ_1 admissible peuvent être écrites sous la forme $S_t = \bar{S}_t + \nu_t$ et $N_t = \bar{N}_t - \nu_t$, où $\bar{S}_t$ et $\bar{N}_t$ sont respectivement les composantes saisonnières et non saisonnières canoniques, et ν_t est un bruit blanc dont la variance est $\bar{\gamma}_1 - \gamma_1$. On peut donc considérer que la décomposition canonique revient à éliminer la plus grande quantité possible de bruit blanc de la composante saisonnière et à le mettre dans la composante irrégulière qui est elle-même comprise dans la composante non saisonnière. Comme aucune raison apparente ne justifie l'inclusion du bruit blanc dans la composante saisonnière, l'emploi de la décomposition canonique semble être une bonne solution. (Watson (1987) propose une approche dans laquelle il n'est pas nécessaire de choisir une décomposition donnée.)

(Par ailleurs, il est bon de noter qu'il est aussi nécessaire de décomposer les effets de la régression déterministe, $\underset{\sim}{X}_t'\underset{\sim}{\beta}$, en des parties saisonnières et non saisonnières. Cette question est traitée dans Bell (1984), mais comme il n'y a aucune raison d'effectuer cette opération différemment dans les modèles ARIMA et dans les modèles de composantes, nous n'aborderons pas cette question dans le présent article.)

Bell et Hillmer (1984) critiquent ceux qui se contentent de prendre les modèles de composantes et de les ajuster à ceux obtenus par modélisation des séries observées, car ils ignorent l'incertitude propre à la décomposition de fixes (de temps en temps) et les valeurs aberrantes font l'objet d'un traitement approfondi. Nous avons exclu quelques séries que Burman et Otto (1988) ont analysées sans publier de résultats, ainsi que des séries de commerce extérieur qu'ils ont analysées depuis qu'elles ont subi des révisions importantes au cours des récentes années, dans le but de corriger certains problèmes majeurs concernant leurs données. Cela nous a laissé 45 séries à analyser, listées dans Bell et Pugh (1990). Ces séries sont largement représentatives des séries désaisonnalisées du bureau de recensement mais, comme elles ne constituent pas un échantillon aléatoire, leur analyse peut être considérée au mieux comme une étude pilote.

Pour une série donnée, nous allons utiliser les mêmes termes de régression avec les modèles ARIMA et les modèles de composantes et nous allons limiter les comparaisons aux modèles dont l'ordre de différenciation est le même. La comparaison de modèles ayant des ordres de différenciation différents présente certains problèmes car les fonctions de vraisemblance des deux modèles sont alors basées sur des données différentes (différenciées). Cette restriction signifie que nous comparerons des modèles ARIMA (2.3) avec d=1 à des BSM conformes à (2.4). Les modèles ARIMA avec d=0 seront comparés à un modèle conforme à (2.4), mais avec T_t conforme à (2.5) avec $\delta=1$. Les modèles dans lesquels la saisonnalité est fixe et d=1 dans la structure ARIMA seront comparés à un modèle de composantes dans lequel la saisonnabilité est fixe (aucun S_t stochastique), et avec T_t à nouveau conforme à (2.5) avec $\delta=1$. Ces deux derniers cas correspondent à des cas particuliers des modèles BSM et GK. Si d=1 et qu'il y a une saisonnalité stochastique dans le modèle ARIMA, nous ne ferons pas de comparaison avec le modèle GK qui utiliserait (2.5) avec $\delta=2$. Comme il s'agit d'un cas spécial de (2.4) avec $\eta=0$, ce modèle GK pourrait au mieux avoir un paramètre supplémentaire de moins et un AIC inférieur de 2 à celui de (2.4). Au pire, il pourrait y avoir un AIC bien supérieur à celui de (2.4), à condition que l'estimation de vraisemblance maximale $\hat{\eta}$ ne soit pas proche de 0 (toutefois si $\hat{\eta} \approx 1$, nous pouvons considérer que (2.4) surdifférencie le modèle GK avec $\delta=1$).

Les modèles ARIMA utilisés et leurs AIC, les BSM ajustés et leurs AIC ainsi que les différences entre les AIC sont donnés dans Bell et Pugh (1990). La table 1 ci-dessous en fournit une récapitulation. Les résultats sont évidents : dans l'ensemble, les critères AIC indiquent une préférence marquée pour les modèles ARIMA, les différences entre les AIC étant importantes (> 8) dans environ la moitié des séries. Les différences entre les AIC de deux séries pour lequel le BSM a été préféré n'étaient que de -2,1 et -2,7.

Table 1 : Comparaison du BSM et du modèle ARIMA

Nombre de séries dans l'intervalle des différences d'AIC	Ordre des différences		
	(1,1)	(0,1)	(1,0)
< -2	2	0	0
-2 à 2	6	1	0
2 à 8	9	2	3
8 à 20	10	3	2
20 à 40	5	0	1
> 40	4	0	0
	36	6	6

(Trois séries figurent deux fois dans la table car elles ont été réajustées avec une saisonnalité fixe après que l'on ait obtenu $\hat{\theta}_{12} \approx 1$.)

La recherche d'explications possibles du mauvais ajustement du BSM nous a conduit à examiner les DAIC et les $\hat{\theta}_{12}$, $\hat{\eta}$ et autres correspondants, mais nous n'avons détecté aucun comportement évident. L'erreur due à la sélection a été considérée comme une explication possible, bien que les modèles ARIMA aient été sélectionnés avec l'approche d'identification habituelle fondée sur des autocorrélations et des autocorrélations partielles et non par recherche du modèle dont l'AIC est minimal dans un ensemble de modèles. Pour étudier l'erreur de sélection, nous avons comparé les AIC des BSM à ceux du "modèle de ligne d'aviation" ARIMA $(0,1,1)\times(0,1,1)_{12}$, qui semble être un choix raisonnable lorsqu'on utilise un modèle ARIMA simple. Bien que le BSM s'ajuste beaucoup mieux que le modèle de ligne d'aviation dans le cas de deux séries (DAIC de -11,7 et de -25,6), les résultats ont été peu différents de ceux figurant à la table 1. Ce n'est peut- être pas tellement surprenant car 15 des modèles ARIMA sélectionnés étaient des modèles de ligne d'aviation et les autres n'étaient guère différents. Le modèle de ligne d'aviation s'est comporté beaucoup mieux que le BSM lorsqu'il a été comparé aux modèles ARIMA sélectionnés, bien que dans quatre séries, le modèle ARIMA sélectionné ait été considéré comme étant supérieur au modèle de ligne d'aviation car l'AIC était supérieur à 20. Cela semble indiquer que l'utilisation pour toutes les séries d'un modèle unique, quel qu'il soit, donne lieu de temps à temps à de mauvais ajustements.

Ce rapport ne serait pas complet sans quelques commentaires au sujet de nos expériences d'ajustement des modèles de composantes. Les résultats présentés ici ont été obtenus à l'aide d'un programme informatique d'ajustement de modèles de séries chronologiques dont les composantes et les termes de régression ARIMA ont récemment été développés par nous-même, par d'autres membres du personnel des séries chronologiques de la division des recherches statistiques du bureau de recensement et par Steven Hillmer de l'université du Kansas.

$$Z_t = S_t + T_t + I_t + V_t \text{ où}$$

$$(1-\alpha_1 B - \ldots - \alpha_p B^p)V_t = \varepsilon_{4t}\ , \ \varepsilon_{4t} \sim \text{iid } N(0,\sigma_4^2) \tag{2.6}$$

avec S_t et I_t comme dans (2.4) et T_t comme dans (2.5). Harvey (1985) étudie également une extension semblable de ses modèles avec des contraintes sur les paramètres autorégressifs, de façon à ce que V_t ait tendance à présenter un comportement cyclique. Il envisage également une formulation ARIMA(2.1) pour V_t.

Les procédures de modélisation de ces modèles de composantes sont plus automatiques que pour les modèles ARIMA et sont traitées dans les références citées. L'estimation se fait à nouveau par vraisemblance maximale, la vraisemblance étant évaluée à l'aide du filtre de Kalman. Comme les modèles sont non stationnaires, l'initialisation du filtre de Kalman présente des difficultés récemment étudiées par Kohn et Ansley (1986) et Bell et Hillmer (1987a). Ces approches produisent une fonction de vraisemblance qui est à nouveau la densité combinée des données différenciées, déterminée maintenant par les modèles de composantes.

Les modèles ARIMA pour les composantes supposent un modèle ARIMA pour le Z_t aggrégatif, ainsi que l'ont noté G. C. Tiao (signalé dans Findley 1983) et Maravall (1985). Dans le cas de (2.4), l'application de $(1-B)^2U(B) = (1-B)(1-B^{12})$ à Z_t donne $(1-B)^2\varepsilon_{1t} + U(B)(1-\eta B)\varepsilon_{2t} + (1-B)(1-B^{12})\varepsilon_{3t}$, ce qui est conforme à un modèle à moyenne mobile d'ordre 13 dont les paramètres sont déterminés par σ_1^2, σ_2^2, σ_3^2, et η. Bien que (2.4) soit en fait équivalent à un modèle ARIMA$(0,1,13)\times(0,1,0)_{12}$ de Z_t, l'ordre de l'opérateur MA hautement régulier et les contraintes sur les paramètres font en sorte qu'il est peu probable qu'une modélisation ARIMA directe de Z_t donne exactement un tel modèle. Donc, il existe des différences possibles entre les approches des modèles ARIMA et des modèles de composantes, bien que Maravall (1985) ait remarqué que certaines valeurs des paramètres de (2.4) donnent un modèle proche du populaire modèle de "ligne d'aviation" ARIMA $(0,1,1)\mathrm{x}(0,1,1)_{12}$ de Box et Jenkins (1976). Pour les séries non saisonnières ou les séries dont la saisonnalité est modélisée comme étant fixe au moyen de la fonction de régression $\underset{\sim}{X}_t'\underset{\sim}{\beta}$, le modèle ARIMA correspondant à (2.4) pour $Z_t = T_t + I_t$ dépend de $(1-\eta B)\varepsilon_{2t} + (1-B)^2 I_t$ qui résulte d'un modèle MA(2) dont les 3 paramètres sont déterminés par σ_2^2, σ_3^2, et η. Nous pourrions facilement obtenir exactement le même modèle ARIMA (0,2,2) par modélisation directe de Z_t. On obtient aussi des résultats semblables avec d'autres modèles de composantes non saisonnières. Bien que les différences possibles entre les modèles ARIMA non saisonniers et les modèles de composantes soient difficiles à identifier, il semble bien plus probable que ces modèles soient réellement identiques dans le cas des données non saisonnières plutôt que dans celui des données saisonnières.

Ceci soulève des questions sur la façon dont le modèle ARIMA et le modèle de composantes diffèrent en pratique et sur celui qui s'ajuste le mieux aux données lorsqu'ils sont différents. Dans une étude préliminaire de cette question, nous allons comparer l'ajustement des modèles ARIMA et des modèles de composantes à un ensemble de séries chronologiques. Comme les modèles que nous désirons comparer ne sont habituellement pas emboités (il n'est pas possible d'obtenir l'un d'entre eux en posant simplement des contraintes sur les paramètres de l'autre), les tests d'hypothèses ou les intervalles de confiance traditionnels seraient difficiles à appliquer. Nous utiliserons le critère AIC d'Akaike (1973), dont la définition est :

$$\mathrm{AIC} = -2\hat{L} + 2m$$

où $\hat{L}$ est la vraisemblance logarithmique maximisée et M est le nombre de paramètres estimés. Le meilleur modèle est celui dont le critère AIC est le plus petit. Pour comparer deux modèles, soient 1 et 2, nous calculons la différence de leur AIC, soit $\mathrm{DAIC} = \mathrm{AIC}_1 - \mathrm{AIC}_2$. Si cette différence est positive, le modèle 2 est supérieur, autrement c'est le modèle 1 qui est le meilleur. Il n'est pas nécessairement facile de déterminer le point où une différence entre les deux modèles peut être considérée comme significative (voir Findley 1988), mais les utilisateurs du critère AIC considèrent fréquemment que des différences de 1 ou 2 le sont. Nous allons considérer que la valeur 2 constitue une limite significative grossière. Une justification rudimentaire consisterait à remarquer que si l'on ajoute un paramètre à un modèle, L ne peut pas diminuer, et que si ce paramètre n'entraîne aucune amélioration de l'ajustement, L reste le même et AIC diminue de 2.

Nous utiliserons le critère AIC pour comparer l'ajustement des modèles ARIMA et des modèles de composantes sur un ensemble de séries chronologiques saisonnières du bureau de recensement analysées par Burman et Otto (1988) à l'aide de modèles ARIMA. (Un grand nombre de séries chronologiques ont déjà été analysées dans Hillmer, Bell et Tiao (1983), mais les données portaient sur un moins grand nombre d'années. Nous avons également étudié une série appelée ENM20 du bureau américain des statistiques sur la main-d'oeuvre qui a été analysée dans Bell et Hillmer 1984.) Ces séries présentent l'avantage d'avoir déjà été associées à des modèles facilement disponibles dans lesquels les termes de régression de la variation temporelle, les effets saisonniers des modèles pour les composantes, ce qui suppose alors qu'un modèle de Z_t peut être ajusté aux données. Nous utiliserons des modèles ARIMA comme base des deux méthodes. Bien que d'autres modèles aient suscité de l'intérêt au cours des récentes années (on pense en particulier aux modèles à mémoire longue, ARCH et non linéaires), les modèles ARIMA et les modèles de composantes ARIMA semblent avoir été les plus populaires et il est donc normal de nous y intéresser initialement.

Les modèles ARIMA que nous utiliserons pour Z_t peuvent être représentés sous la forme (Box et Jenkins 1976) :

$$\phi(B)(1-B)^d(1-B^{12})Z_t = \theta(B)(1-\theta_{12}B^{12})a_t \tag{2.3}$$

où B est l'opérateur de retard ($BZ_t = Z_{t-1}$), $d \geq 0$ (si d=0, $(1-B)^d=1$), $\phi(B) = 1-\phi_1B-\ldots-\phi_pB^p$ et où $\theta(B) = 1-\theta_1B-\ldots-\theta_qB^q$ sont des opérateurs AR et MA d'ordre faible (habituellement p et $q \leq 3$) et a_t est un bruit blanc (les a_t, $1 \leq t \leq n$, sont indépendants, identiquement distribués). Ce modèle est destiné aux données saisonnières mensuelles; les modifications à apporter aux données pour les autres périodes saisonnières (p. ex. trimestrielles) sont évidentes et les facteurs $1-B^{12}$ et $1-\theta_{12}B^{12}$ sont éliminés dans le cas des données non saisonnières. Nous pourrions également inclure un opérateur autorégressif saisonnier dans (2.3), mais nous le faisons rarement. Si $\theta_{12} = 1$, nous pouvons annuler le facteur $1-B^{12}$ des deux côtés de (2.3) et ajouter à $\underset{\sim}{X}_t$ des variables de moyennes saisonnières (Abraham et Box 1978, Bell 1987). L'identification, l'estimation et la vérification du diagnostic de ces modèles s'effectuent à l'aide de procédures qui sont maintenant devenues bien établies - se reporter à Box et Jenkins (1976) pour les modèles ARIMA purs, à Bell et Hillmer (1983) et à Findley et al. (1988) pour les modèles comprenant des termes de régression. L'estimation s'effectue par vraisemblance maximale, la fonction de vraisemblance étant définie comme étant la densité combinée des données différenciées $(1-B)^d(1-B^{12})Y_t$, $t=d+13,\ldots,n$.

Les modèles de composantes spécifient des modèles ARIMA simples pour les composantes dans (2.2). Le modèle structurel de base (BSM) de Harvey et Todd (1983) peut s'écrire sous la forme :

$$\begin{gathered} Z_t = S_t + T_t + I_t \text{ où} \\ U(B)\ S_t = \varepsilon_{1t}\ , \quad \varepsilon_{1t} \sim \text{iid } N(0,\sigma_1^2), \\ (1-B)^2T_t = (1-\eta B)\varepsilon_{2t}\ , \quad \varepsilon_{2t} \sim \text{iid } N(0,\sigma_2^2), \\ I_t \sim \text{iid } N(0,\sigma_3^2) \end{gathered} \tag{2.4}$$

où $U(B) = 1 + B \ldots + B^{11}$ est la somme d'une série calculée sur 12 mois consécutifs. En fait, les auteurs commencent par T_t après une marche aléatoire avec dérive "stochastique", dans laquelle la dérive suit également une marche aléatoire; ceci conduit au modèle (0,2,1) pour T_t dans (2.4) avec la contrainte $\eta \geq 0$. Bien que nous remettions pas en vigueur cette contrainte, elle n'en est pas moins facilement satisfaite dans toutes les séries données en exemple dans cet article. Si la dérive "stochastique" a une variance d'innovation nulle (en fait, il s'agit d'une constante), $\eta = 1$, et le modèle de T_t peut être réduit à $(1-B)T_t = \beta_0 + \varepsilon_{1t}$. Nous pouvons tenir compte de β_0 en ajoutant la variable de tendance temporelle t à $\underset{\sim}{X}_t$. Si $\sigma_1^2 = 0$, S_t devient fixe et peut être traité dans $\underset{\sim}{X}_t$ avec des variables appropriées semblables à celles indiquées lorsque $\theta_{12} = 1$ dans le modèle ARIMA (2.3).

Gersch et Kitagawa (1983) (voir aussi Kitagawa et Gerch 1984) étudient des modeles semblables à (2.4), mais avec T_t conforme au modèle suivant :

$$(1-B)^\delta T_t = \varepsilon_{2t}\ , \ \delta = 1,2 \text{ ou } 3. \tag{2.5}$$

Le modèle GK est donc (2.4) avec T_t conforme au modèle (2.5). On remarquera que le modèle GK avec $\delta = 2$ est équivalent au modèle BSM avec $\eta = 0$, tandis que le BSM avec $\eta=1$ est équivalent au modèle GK avec $\delta=1$ muni d'une constante de tendance. Akaike (1980) propose des modèles semblables, mais avec S_t conforme à un modèle qui semble maintenant peu attrayant.

Gersch et Kitagawa ont développé leur modèle en y ajoutant une composante autorégressive stationnaire. Ce nouveau modèle peut s'écrire sous la forme :

populaire de "ligne d'aviation" (airline) ARIMA $(0,1,1)\times(0,1,1)_{12}$ de Box et Jenkins (1976), en montrant que les autocorrélations des séries différenciées pouvaient être semblables pour les deux modèles (suivant les valeurs des paramètres). Ceci a fait entrevoir la possibilité importante que le BMS et certains modèles ARIMA pouvaient être pratiquement les mêmes dans le cas de certaines séries. Carlin et Dempster (1989), dans une analyse détaillée des deux séries, n'ont détecté que des petites différences entre les désaisonalisations canoniques ARIMA et celles d'un modèle de composantes à moyenne mobile intégré fractionnellement (FRIMA), et des différences plus importantes dans la comparaison de la désaisonnalisation FRIMA et de la désaisonnalisation X-11 normalement utilisées pour d'autres séries.

La documentation semble laisser sans réponse deux questions importantes, soient : (1) Les modèles ARIMA ou les modèles de composantes s'ajustent-ils mieux aux données réelles ou serait-il possible de les distinguer à l'aide des données disponibles. (2) Jusqu'à quel point les résultats du modèle de composantes et du modèle ARIMA sont-ils différents dans les applications pratiques? La première question vise la signification statistique et la seconde la signification pratique. Ces questions sont toutes deux très empiriques et l'objet principal de cet article est de les traiter empiriquement. A la section 2, nous décrivons les modèles que nous allons traiter en détail. Nous utilisons le critère AIC de Akaike (1973) pour comparer l'ajustement des modèles ARIMA et des BSM à un ensemble de 45 séries chronologiques saisonnières. En général, le critère AIC semble largement favoriser les modèles ARIMA.

La section 3 traite de la désaisonnalisation. Bell et Hillmer (1984) ont remarqué que les modèles de composantes avaient ignoré l'incertitude inhérente des décompositions saisonnières et non saisonnières qui caractérise n'importe quel modèle ajusté. Pour étudier cette question, nous avons considéré l'intervalle des décompositions admissibles caractérisant un modèle de composantes donné et nous avons présenté une "décomposition canonique" pour des modèles de composantes analogue à celle proposée pour des modèles ARIMA par Burman (1980) et Hillmer et Tiao (1982). La décomposition canonique s'est révélée triviale à obtenir et très facile à utiliser dans l'extraction des signaux de la désaisonnalisation. En outre, elle s'est également révélée très proche du modèle de composantes ajusté initialement aux séries étudiées dans le présent article, ce qui semble indiquer que la désaisonnalisation dans les modèles de composantes canoniques et dans le modèle initial pourraient souvent être pratiquement identiques. Nous avons alors comparé les corrections des variations saisonnières du BSM et du modèle ARIMA dans le cas de deux séries et nous avons constaté qu'il y avait des différences négligeables dans les estimations des points d'extraction des signaux et des différences relativement grandes dans les variances d'extraction des signaux, bien que celles-ci semblent être toutes petites dans l'absolu.

A la section 4, nous avons étudié les effets de l'utilisation des modèles ARIMA et des modèles de composantes sur l'application des techniques d'extraction des signaux des séries chronologiques dans le but d'estimer des enquêtes répétées. Cette idée avait été premièrement suggérée par Scott et Smith (1974) et par Scott, Smith et Jones (1977), mais elle a fait l'objet d'études poussées plus récentes en raison de l'évolution des méthodes de calcul et des méthodes théoriques d'estimation et d'extraction des signaux des modèles des séries chronologiques non stationnaires. Pour chacune des deux séries, nous avons constaté que les estimations des points d'extraction des signaux obtenues avec les modèles ARIMA et les BSM étaient assez proches, mais que, dans le cas d'une série, les variances d'extraction des signaux étaient assez différentes. Enfin, à la section 5, nous avons tiré quelques conclusions préliminaires.

2. MODÈLES ARIMA ET MODÈLES DE COMPOSANTES

Soit Y_t, $t = 1,\ldots,n$, les observations d'une série chronologique qui, dans bien des cas, est le logarithme d'une certaine série chronologique originale. Posons

$$Y_t = \underset{\sim}{X}_t'\underset{\sim}{\beta} + Z_t \tag{2.1}$$

où $\underset{\sim}{X}_t'\underset{\sim}{\beta}$ est une fonction moyenne de régression linéaire, $\underset{\sim}{X}_t$ est le vecteur des variables de régression au temps t, $\underset{\sim}{\beta}$, est le vecteur des paramètres de régression et Z_t est la partie stochastique (moyenne nulle) de Y_t. Les variables de régression utilisées serviront à tenir compte des constantes de tendance, de la variation temporelle, des effets saisonniers fixes et des effets des valeurs aberrantes (Findley *et al.* 1988). Nous étudierons des décompositions de Z_t telles que :

$$Z_t = S_t + N_t = S_t + T_t + I_t \tag{2.2}$$

où S_t est une composante saisonnière (stochastique) et N_t une composante non saisonnière (stochastique) qui peut être décomposée en une composante de tendance T_t et en une composante irrégulière I_t. Si Y_t est le logarithme de la série chronologique étudiée, (2.1) et (2.2) supposent qu'il existe des décompositions multiplicatives de la série chronologique originale. Une approche à l'analyse des composantes des séries chronologiques consiste à modéliser directement Z_t, puis à faire des hypothèses permettant d'arriver à partir de ce modèle à des définitions et à des modèle de composantes. L'autre méthode consiste à spécifier directement

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

AUTRES APPROCHES D'ANALYSE DES ÉLÉMENTS DES SÉRIES CHRONOLOGIQUES

W.R. Bell et M.G. Pugh[1]

RÉSUMÉ

La documentation récente sur l'analyse des séries chronologiques propose diverses approches fondées sur une structure particulière des composantes. Ces approches sont différentes de l'approche de modélisation ARIMA (moyenne mobile, intégrée, autorégressive) bien connue, la plus populaire étant peut-être l'approche par "modélisation structurelle" de Harvey et autres qui utilise une structure explicite pour les composantes. Malgré le volume considérable de recherches effectuées sur ces modèles, très peu de travaux semblent avoir été entrepris pour comparer les résultats des autres approches. Nous sommes alors amenés à se poser des questions sur l'ajustement comparatif de ces modèles et de l'effet du choix du modèle sur des applications telles la désaisonnalisation (fondée sur un modèle) et l'utilisation des modèles de séries chronologiques dans l'estimation des enquêtes répétées. Comme il s'agit de questions empiriques, nous avons essayé de les traiter dans le présent article en comparant les résultats de l'application de ces modèles à certaines séries chronologiques du bureau du recensement.

MOTS CLÉS: Modèle ARIMA; modèle des composantes; AIC; désaisonnalisation; estimation des enquêtes répétitives.

1. INTRODUCTION

L'analyse des composantes des séries chronologiques a connu une longue évolution (traitée dans Nerlove, Grether et Carvalho 1979) depuis les travaux d'astronomie, de météorologie et d'économie effectués du 17^{e} au 19^{e} siècle jusqu'aux premières analyses saisonnières par Buys-Ballot (1847). Des méthodes empiriques de désaisonnalisation ont été créées au début de notre siècle et ont abouti en 1967 à l'élaboration de la méthode bien connue X-11. Bell et Hillmer (1984) ont expliqué que ces méthodes ont précédé les modèles des séries chronologiques saisonnières qui se sont largement répandus et sont devenus calculables au cours des vingt dernières années.

Cet intérêt de longue date à l'égard des composantes des séries chronologiques a eu des influences importantes sur la modélisation des séries chronologiques; en particulier, il a conduit à deux approches assez différentes de modélisation et de désaisonnalisation à l'aide de modèles. Plusieurs approches de désaisonnalisation ont été élaborées avec les modèles à moyenne mobile, intégrés, autorégressifs (ARIMA, Box et Jenkins 1976). A notre avis, la meilleure est l'approche "canonique" de Burman (1980) et de Hillmer et Tiao (1982). Une approche différente de "modélisation des composantes" a été élaborée. Elle utilise des modèles ARIMA simples pour les composantes saisonnières, tendancielles, régulières et autres. Nerlove, Grether et Carvalho (1979) ont proposé une approche quelque peu différente, (sans grand succès cependant), probablement parce que leurs modèles de composantes ARIMA sont trop souples pour même permettre l'identification de la structure du modèle (Hotta 1989), et qui est maintenant considérée inadéquate à cause de son traitement de la non-stationarité (par extraction de la tendance polynomiale). Des exemples marquants de cette approche sont les travaux d'Akaike (1980), de Gersch et Kitagawa (1983), de Kitagawa et Gersch (1984), de Harvey et Todd (1983) et de Harvey (1985).

Bien que des travaux de développement considérables aient été effectués sur les deux approches de modélisation, il est étonnant de constater qu'il y a très peu de documentation sur la comparaison de leurs résultats. Harvey et Todd (1983) ont comparé les prévisions obtenues avec leur "modèle structurel de base" (BSM) et celles obtenues avec les modèles ARIMA ajustés par Prothero et Wallis (1976) à six séries chronologiques trimestrielles de macroéconomie. Leurs résultats étaient assez peu concluants. D'ailleurs certains des modèles ARIMA utilisés avaient une forme inhabituelle et étaient, en particulier, caractérisés par des opérateurs saisonniers correspondant à des retards importants. (En toute objectivité, les travaux de Prothero et Wallis (1976) sur la modélisation saisonnière ARIMA n'étaient pas très avancés avant que des raffinements tels la vraisemblance maximale exacte et le traitement des valeurs aberrantes ne deviennent disponibles.) Harvey (1985) a élargi le BSM en développant des modèles de composantes afin d'expliquer le comportement cyclique (séries non saisonnières) et a traité jusqu'à un certain point leurs relations avec les modèles ARIMA. Maravall (1985) a observé que le BSM pouvait fournir un modèle général proche du modèle populaire de "ligne d'aviation" (airline) ARIMA $(0,1,1)\times(0,1,1)_{12}$ de Box et Jenkins (1976), en montrant que

modèles ARIMA. Maravall (1985) a observé que le BSM pouvait fournir un modèle général proche du modèle

[1] W.R. Bell, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A. M.G. Pugh, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

SECTION 4

DÉVELOPPEMENTS DANS L'ANALYSE DE DONNÉES DES SÉRIES CHRONOLOGIQUES

NOMBRE DE DÉCÈS ATTRIBUABLES
AU CANCER DU POUMON
Nombre d'hommes

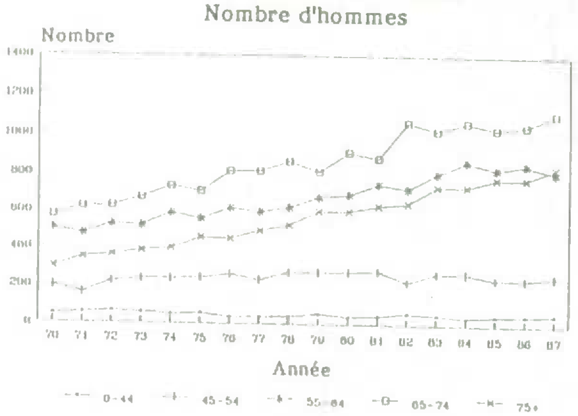
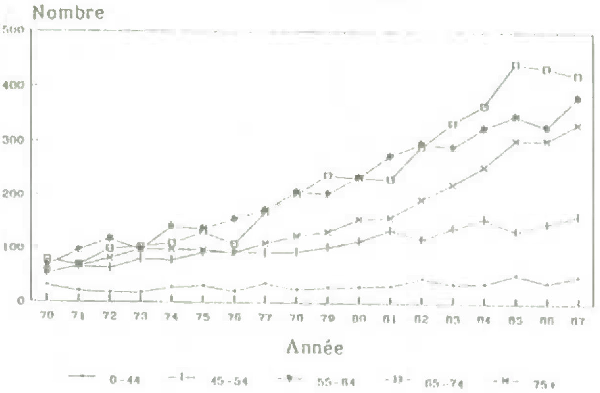


Figure 1

NOMBRE DE DÉCÈS ATTRIBUABLES
AU CANCER DU POUMON
Nombre de femmes



ONTARIO

Figure 2

VALEURS THETA TRANSVERSALES

1970-1987

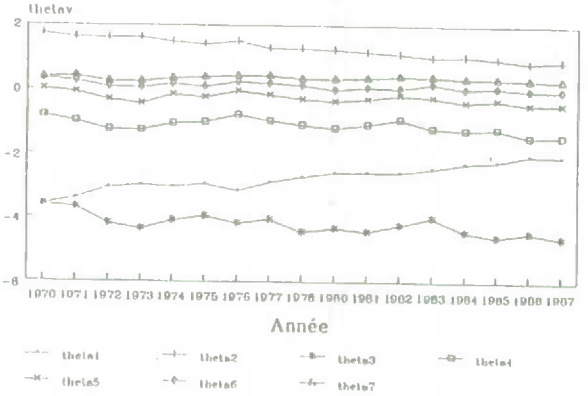


Figure 3

VALEURS DES PREMIÈRES
DIFFÉRENCES DES THETA
1971-1987

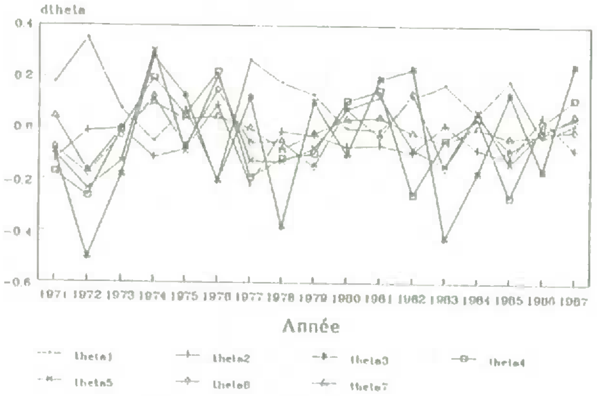


Figure 4

NOMBRE DE DÉCÈS ATTRIBUABLES
AU CANCER DU POUMON
Hommes, 65-74 - pas d'écart

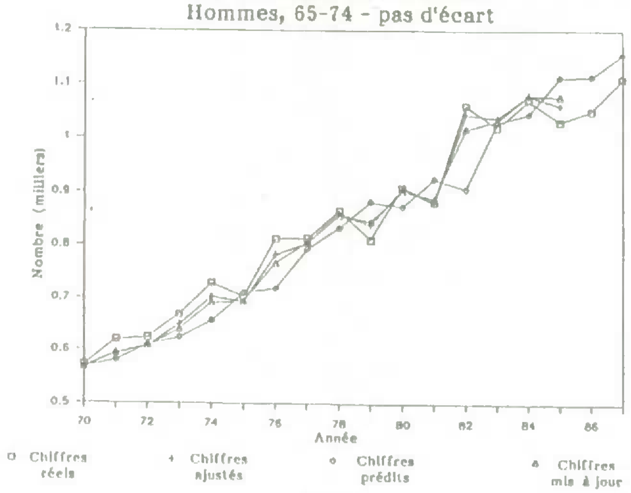


Figure 5

NOMBRE DE DÉCÈS ATTRIBUABLES
AU CANCER DU POUMON
Femmes, 55-64 - pas d'écart

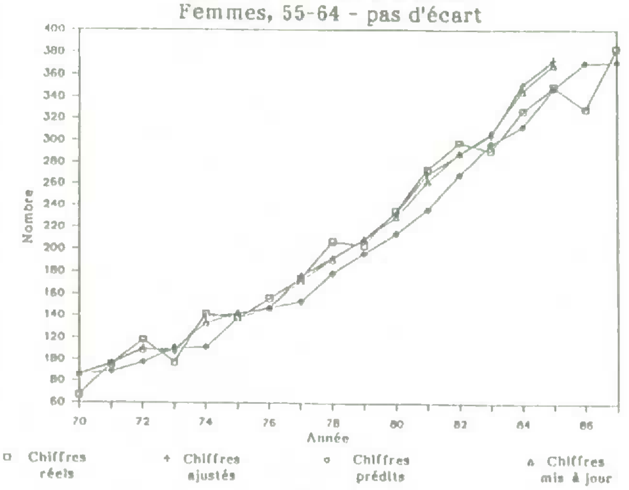


Figure 6

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, *33*, 133-158.

Kohn, R. and Ansley, C.F. (1989). A fast algorithm for signal extraction, influence, and cross-validation in state space models. *Biometrika*, 65-79.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika*, *74*, 247-258.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, *11*, 59-67.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* (2nd ed.). London: Chapman and Hall.

Pfeffermann, D. (1989). Estimation and seasonal adjustment of population means using data from repeated surveys. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from a cluster sample. *Jour. Amer. Statist. Assoc.*, *76*, 681-689.

Preisler, H. (1989). Analysis of a toxicological experiment using a generalized linear model with nested random effects. *Int. Statist. Rev.* *57*, 145-159.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Ann. Statist.*, *12*, 46-60.

Smith, T.M.F. and Brunsdon, T.M. (1989). The time series analysis of compositional data. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Stiratelli, R., Laird, N.M., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, *40*, 961-972.

Stram, D.O., Wei, L.J., and Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Jour. Amer. Statist. Assoc.*, *83*, 631-637.

Swamy, P.A.V.B. (1970). Efficient inferences in a random coefficient regression model. *Econometrica*, *38*, 311-323.

West, M., Harrison, J.P. and Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *Jour. Amer. Statist. Assoc.*, *80*, 73-96.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika*, *75*, 621-629.

Zeger, S.L., Liang, K.-Y. and Self, S.G. (1985). The analysis of binary longitudinal data with time independent covariates. *Biometrika*, *72*, 31-38.

Zeger, S.L., and Qaqish, B. (1988). Markov regression models for time series; a quasi-likelihood approach. *Biometrics*, 44,1019-1031.

Zehnwirth, B. (1988). A generalization of the Kalman filter for models with state-dependent observation variance. *Jour. Amer. Statist. Assoc.*, *83*, 164-167.

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregation bias. *Jour. Amer. Statist. Assoc.* *57*, 348-368.

extensions du MLGEÉ justifieraient des études. Par exemple, l'inclusion des effets saisonniers pour les séries mensuelles ou trimestrielles aussi bien que d'effets d'intervention dans l'équation de transition du MLGEÉ serait souhaitable. Il est possible de modifier le cadre actuel du MLGEÉ afin d'inclure des effets saisonniers ou d'intervention non stochastiques. Cependant, la présence d'effets stochastiques devrait faire l'objet d'études plus poussées. Par ailleurs, pour les séries chronologiques résultant d'enquêtes complexes, il serait important d'étudier l'impact de plans complexes sur l'inférence à propos des paramètres de modèle analogues aux corrections de Rao et Scott (1984) pour l'analyse de données transversale. Dans le cas des enquêtes par panel se pose le problème supplémentaire des erreurs d'enquêtes corrélées dans l'équation de mesure, en raison du chevauchement des unités entre des points successifs dans le temps, telles qu'elles ont été étudiées par Binder et Dick (1989) et Pfeffermann (1989) pour la modélisation ARMM des erreurs d'enquête dans le contexte des modèles linéaires à espace d'états.

REMERCIEMENTS

La recherche du premier auteur, qui eut lieu à l'université Carleton alors qu'il était professeur adjoint de recherches, a été financée en partie par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs tiennent également à remercier D. Binder et J.N.K. Rao pour leurs commentaires utiles et en particulier D. Pfeffermann pour ses remarques judicieuses et plusieurs discussions utiles. Ils tiennent aussi à remercier Christine Larabie, Carmen Lacroix et en particulier Judy Clarke pour la préparation du manuscrit dans des délais extrêmement courts.

RÉFÉRENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: John Wiley.

Binder, D.A. and Dick, P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology, 15,* 29-45.

Box, G.E.P. and Jenkins, G.M. (1970). *Time series analysis: forecasting and control.* San Francisco: Holden-Day.

Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics, 42,* 693-734.

Cox, D.R. (1981). Statistical analysis of time series, some recent developments (with discussion). *Scand. Jour. Statist., 8,* 93-115.

Fuller, W.A. (1976). *Introduction to Statistical Time Series.* New York: John Wiley.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics, 25,* 489-504.

Harvey, A.C. (1981). *Time Series Models,* Oxford: Philip Allan, and New York: John Wiley.

Harvey, A.C. (1984). A unified view of statistical forecasting procedures (with discussion). *Jour. Forecasting, 3,* 245-275.

Harvey, A.C. and Peters, S. (1984). Estimation procedures for structural time series models. *London School of Economics, Discussion Paper* A. 44.

Harvey, A.C. and Durbin, J. (1986). The effects of seat belt legislation on British road casualties: A case study in statistical time series modelling (with discussion). *Jour. Roy. Statist. Soc. A, 149,* 187-227.

Harvey, A.C. and Fernandes, C. (1989). Time series models for counts or qualitative observations. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

Kalbfleisch, J.D. and Lawless, J.F. (1984). Least-squares estimation of transition probabilities from aggregate data. *Can. Jour. Statist. 12,* 169-182.

Kalbfleisch, J.D. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Jour. Amer. Statist. Assoc., 80,* 863-871.

Kaufmann, H. (1987). Regression methods for non-stationary categorical time series: Asymptotic estimation theory. *Ann. Statist., 17,* 79-98.

Kitagawa, G. (1987). Non-Gaussian state space modelling for non-stationary time series (with discussion). *Jour. Amer. Statist. Assoc., 82,* 1032-1063.

Tableau 1: Écarts-types estimés des résidus de prédiction un pas en avant normalisés

i	1	2	3	4	5	6	7	8	9	10
$ÉT(r_{it})$	.76	.86	.70	.99	.50	.88	.69	.68	.99	.72

On ne relève aucun indice de surdispersion, parce que tous les $ÉT(r_{it})$ sont inférieurs à un. Un graphique des r_{it} en fonction du temps n'a révélé aucune mauvaise spécification. Les tests prédictifs post-échantillon pour les projections un pas et deux pas en avant ont été effectués pour les deux dernières années (1986 et 1987). On a obtenu les valeurs de X^2 suivantes

$$X_1^2 = 9.40 \quad , \quad X_2^2 = 1.14 \ , \tag{5.5}$$

qui, lorsque confrontées à une distribution X_{10}^2, étaient clairement non significatives. La SCE a été calculée comme étant égale à 167.5 avec une estimation du paramètre de surdispersion de 1.046. Là encore, il n'a pas semblé y avoir de surdispersion. On a obtenu des valeurs de REQMEW_1, $\text{REQMEP}_1(1)$ et $\text{REQMEP}_1(2)$ de 31.1, 28.1 et 17.6 respectivement. Les valeurs correspondantes pour les modèles simples étaient 84.7, 26.4 et 56.1. Le MLGEÉ semble apporter une amélioration considérable par rapport au modèle simple. On pourra remarquer que lors du calcul de REQMEP(1), il n'y a qu'un point dans le temps (1986) pour lequel on a fait des prédictions en utilisant les données allant jusqu'en 1985.

Les figures 5 et 6 présentent le graphique des échantillons du nombre de décès réels, ajustés transversalement, prédits un pas en avant (deux pas pour le dernier point) et filtrés (ou mis à jour), pour les hommes du groupe d'âge 65-74 et pour les femmes du groupe d'âge 56-64 respectivement. Le tableau 2 donne un résumé des nombres prédits ainsi que les nombres réels pour tous les dix groupes pour 1986 et 1987. Les valeurs de la REQME des nombres prédits figurent entre parenthèses.

Tableau 2: Nombre de cas prédits (P) et réels (R) de décès dus au cancer du poumon en Ontario

Groupe_		Hommes					Femmes				
d'âge:		1	2	3	4	5	1	2	3	4	5
1986	P:	58 (14)	246 (36)	818 (64)	1114 (74)	799 (57)	46 (14)	143 (20)	370 (42)	424 (53)	317 (55)
	R:	51	242	851	1050	775	38	150	329	435	304
1987	P:	59 (18)	250 (48)	823 (83)	1155 (99)	833 (74)	47 (18)	146 (25)	371 (55)	441 (74)	329 (76)
	R:	56	256	810	1110	835	50	164	383	422	334

6. DISCUSSION

On a montré que si le nombre de points dans le temps et le nombre d'observations à chaque point dans le temps sont assez grands, on peut transformer de façon appropriée les données provenant de séries chronologiques non normales et non stationnaires de modèles possiblement non linéaires, pour l'application des techniques de modélisation linéaire à espace d'états. Les estimations de paramètres transversales et convergentes $\{\hat{\underline{\theta}}_t^C\}$ peuvent servir à spécifier approximativement la matrice des variances et covariances W de l'équation de transition lorsque T est grand et lorsque W_t est supposée invariable dans le temps. On notera que si la matrice de transition G_t contient certains paramètres inconnus, ils peuvent aussi être estimés de façon convergente en utilisant l'estimateur Aitken à deux pas de Zellner (1962) présenté dans le contexte d'équations de régression apparemment non reliées. On a également montré que lorsque T n'est pas grand, les inférences à propos de $\underline{\theta}_t$ restent robustes devant une mauvaise spécification de W, à condition que la fonction moyenne soit correctement spécifiée.

Comme la méthode du filtre de Kalman (Harvey, 1984) peut de routine régler les problèmes de données manquantes lorsque l'on suppose que les observations sont également espacées, la méthode du MLGEÉ proposée peut également être appliquée à ces situations. Toutefois, il y a certaines directions pour lesquelles des

Pour l'ajustement du MLGEÉ, nous devons d'abord spécifier le comportement transversal au temps t. Il y a dix groupes, donc m=10. Pour le groupe i le nombre y_{it}, pour chaque i, suit par hypothèse une distribution de Poisson au sens large avec moyenne μ_{it} égale à $n_{it}\ \lambda_{it}$, où n_{it} est la taille connue de la population. On a considéré un modèle log-linéaire pour les taux $\{\lambda_{it},\ i=1,\ \ldots,\ 10\}$. Un modèle à effets de lignes pour les données ordinales lorsque la ligne et la colonne se rapportent respectivement au sexe et à l'âge (Agresti, 1984, p. 84) a donné un ajustement raisonnable transversalement pour presque tous les points dans le temps. On a donc choisi le modèle transversal avec des scores convenables pour les catégories d'âge comme suit

$$\log \lambda_{it} = \underset{\sim}{f}_i' \underset{\sim}{\theta}_t,\ i=1,\ 2,\ \ldots,\ 10 \tag{5.1a}$$

où

$$F = \begin{bmatrix} f_1' \\ f_2' \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ f_{10}' \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & -2 \\ 1 & 1 & 0 & 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 2 \\ \hdashline 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.1b}$$

Il y a un total de sept effets $\theta_1,\ \theta_2, \ldots, \theta_7$: un pour la constante, un pour le sexe, quatre pour l'âge et un pour l'interaction âge-sexe. Notons qu'ici la matrice des variables auxiliaires, F_t, est invariable dans le temps . La matrice des variances et covariances $U_t(\underset{\sim}{\theta}_t)$ est évaluée à l'estimation $\hat{\underset{\sim}{\theta}}{}_t^C$ comme suit

$$U_t(\hat{\underset{\sim}{\theta}}{}_t^C) = \text{diag}\ (\hat{\underset{\sim}{\theta}}{}_t^C)^{-1} \tag{5.2}$$

On obtient la forme de (5.2) facilement en utilisant la relation variance-moyenne de Poisson et la fonction de lien log.

Ensuite, pour spécifier le comportement longitudinal dans le MLGEÉ, on a étudié les graphiques de $\hat{\theta}_{it}^C$ et de $\hat{\theta}_{it}^C - \hat{\theta}_{it-1}^C$ en fonction du temps pour chaque $i=1,\ \ldots,\ 10$; voir les figures 3 et 4. La série des premières différences de $\hat{\theta}_{it}^C$ semble assez aléatoire autour d'une moyenne zéro, hormis une légère déviation de la série des $\hat{\theta}_{1t}^C$. On pourrait faire la régression de $\hat{\theta}_{it}^C$ sur $\hat{\theta}_{it-1}^C$ pour chaque i et vérifier le comportement aléatoire des résidus au lieu de celui des premières différences . Toutefois, pour des considérations d'illustration et de simplicité, on a choisi une marche aléatoire avec un modèle sans déviation pour représenter l'équation de transition, c.-à-d.

$$\underset{\sim}{\theta}_t = \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\xi}_t \tag{5.3}$$

On a donc finalement supposé que la matrice de transition G_t était également invariable dans le temps et on l'a posée égale à I. On a estimé la covariance de $\underset{\sim}{\xi}_t$ par

$$\hat{W} = \frac{1}{T-1} \sum_{t=2}^{T} (\hat{\underset{\sim}{\theta}}{}_t^C - \hat{\underset{\sim}{\theta}}{}_{t-1}^C)(\hat{\underset{\sim}{\theta}}{}_t^C - \hat{\underset{\sim}{\theta}}{}_{t-1}^C)' \tag{5.4}$$

Après avoir spécifié le MLGEÉ, on a ajusté le modèle en utilisant l'algorithme des MCPIF déjà mentionné à la section 3.2. Les écarts-types de l'échantillon des résidus un pas en avant normalisés r_{it}, $i=1, \ldots, 10$, qui furent obtenus sont

Considérons maintenant la situation où W_t inconnue = W, mais T n'est pas grand. Dans ce cas, $\hat{W}$ ne sera pas une bonne spécification de W, parce qu'il est peu probable qu'elle soit dans le voisinage de W. Bien que la propriété MPLN des estimations ne tienne plus, la proposition suivante montre que les estimations $\underset{\sim}{\hat{\theta}}{}^P_T$ et $\underset{\sim}{\hat{\theta}}{}^L_T$ restent convergentes pour n_t grand. Cette propriété d'inférence robuste concernant $\underset{\sim}{\theta}_t$ est semblable à celle obtenue par Zeger (1988) pour les modèles de régression de séries chronologiques de comptes.

Proposition 4.3 Supposons que la fonction moyenne est correctement spécifiée en termes de F_t et de G_t, mais que W peut être mal spécifiée par $\hat{W}$ lorsque T n'est pas grand. Alors, pour n_t grand et pour $t \geq 2$, les moyennes asymptotiques des distributions de $\underset{\sim}{\hat{\theta}}{}^P_{t|t-1} - \underset{\sim}{\theta}_t$ et de $\underset{\sim}{\hat{\theta}}{}^L_{t|t} - \underset{\sim}{\theta}_t$ restent les mêmes, c.-à-d. zéro, mais leurs EQM deviennent $C^*_{t|t-1}$ et C^*_t respectivement lorsque, pour $t \geq 2$,

$$C^*_{t|t-1} = G_t\, C^*_{t-1}\, G'_t + W, \tag{4.4}$$

$$C^*_t = (I - \hat{K}_t F_t)\, \hat{C}_{t|t-1}(\hat{C}^{-1}_{t|t-1}\, C^*_{t|t-1}\, \hat{C}^{-1}_{t|t-1} + F'_t\, U^{-1}_t F_t)\, \hat{C}_{t|t-1}(I - \hat{K}_t F_t)' \quad , \tag{4.5}$$

et "^" indique que $\hat{W}$ est substituée à W. Dans (4.4) ci- dessus, pour t=1, C^*_1 est la même que C_1 définie plus tôt par $(F'_1\ U^{-1}_1\ F_1)^{-1}$.

La démonstration de la proposition ci-dessus se présente comme suit. À la suite de Zehnwirth (1988), on peut exprimer l'estimateur $\underset{\sim}{\hat{\theta}}{}^L_t$ comme une combinaison linéaire de $\underset{\sim}{\hat{\theta}}{}^P_{t|t-1}$ et de l'estimateur $\underset{\sim}{\hat{\theta}}{}^C_t$ ou $(F'_t U^{-1}_t\ F_t)^{-1}\ F'_t\ U^{-1}_t\ \underset{\sim}{z}_t$, c.-à-d.

$$\underset{\sim}{\hat{\theta}}{}^L_t = (I - \hat{\Lambda})\ \underset{\sim}{\hat{\theta}}{}^P_{t|t-1} + \hat{\Lambda}\ \hat{\theta}^C_t \tag{4.6}$$

où

$$\Lambda = (I - K_t F_t)\ C_{t|t-1}\ F'_t\ U^{-1}_t\ F_t, \tag{4.7}$$

et $\hat{\Lambda}$ correspond à Λ lorsque $\hat{W}$ est substituée à W. À mesure que $n_t \to \infty$, la convergence de $\underset{\sim}{\hat{\theta}}{}^L_t$ suit facilement par induction, en commençant à t=2, 3, ... et ainsi de suite. Pour obtenir l'expression (4.4) pour C^*_t, écrivons une autre expression équivalente pour (4.6) sous la forme

$$\underset{\sim}{\hat{\theta}}{}^L_t = \underset{\sim}{\hat{\theta}}{}^P_{t|t-1} + \hat{K}_t(\underset{\sim}{z}_t - F_t\ \underset{\sim}{\hat{\theta}}{}^P_{t|t-1}) \tag{4.8a}$$

$$= (I - \hat{K}_t F_t)\ \hat{C}_{t|t-1}(\hat{C}^{-1}_{t|t-1}\ \underset{\sim}{\hat{\theta}}{}^P_{t|t-1} + F'_t\ U^{-1}_t\ \underset{\sim}{z}_t), \tag{4.8b}$$

en raison de l'identité

$$(I - K_t F_t)\ C_{t|t-1}\ F'_t U^{-1}_t = K_t\ . \tag{4.9}$$

Les résultats souhaités (4.4) et (4.5) découlent immédiatement de (4.8b). On remarquera que lorsque $\hat{W} \approx W$, alors compte tenu de (4.9), C^*_t de (4.5) se réduit approximativement à C_t ou $(I - K_t F_t)\ C_{t|t-1}$, comme on pouvait s'y attendre.

5. APPLICATION AUX SÉRIES DU NOMBRE DE DÉCÈS ATTRIBUABLES AU CANCER

À des fins d'illustration du MLGEÉ, on a analysé les données sur le nombre de décès annuels attribuables au cancer du poumon en Ontario pour la période de 1970-1987, regroupées selon le sexe et l'âge. On a retenu cinq groupes d'âge: 1 = 0-44, 2 = 45-54, 3 = 55-64, 4 = 65-74 et 5 = 75+. Les figures 1 et 2 présentent les dix séries chronologiques du nombre de décès, classées selon le sexe. Nous avons utilisé les données de 1970-1985 (c.-à-d. 16 points dans le temps) pour ajuster le modèle, et les données des deux points suivants (1986, 1987) pour les diagnostics post-échantillon.

Les prédicteurs lissés de $\underset{\sim}{\theta}_t$ ou de $\underset{\sim}{y}_t$ pour tout point $t<T$ étant donné toutes les observations $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_T$ peuvent être calculés grâce à l'algorithme donné par Harvey (1981, p. 115) ou par un algorithme rapide de Kohn et Ansley (1989).

4. QUELQUES RÉSULTATS THÉORIQUES

Supposons que n_t est grand pour chaque t, de sorte que les $U_t(\hat{\underset{\sim}{\theta}}{}_t^C)$ fournissent effectivement des matrices de variances et covariances approximatives pour les $\underset{\sim}{z}_t$ conditionnellement sur les θ_t. On suppose que la matrice des variances et covariances W_t est connue pour les propositions 4.1 et 4.2. Par contre, dans la proposition 4.3, nous étudions l'effet d'une mauvaise spécification de W sur les estimations de $\underset{\sim}{\theta}_t$ lorsque T n'est pas grand, c.-à-d. lorsque $\hat{W}$ est instable. La proposition suivante donne la distribution asymptotique de X^2_τ définie plus tôt par (3.10).

Proposition 4.1 Supposons que les données sont groupées en m-vecteurs $\underset{\sim}{y}_t$ qui, pour n_t grand, sont asymptotiquement normaux sous une application appropriée du TLC. Alors

$$X^2_\tau \doteq \chi^2_m \tag{4.1}$$

Pour cela, remarquons que

$$\underset{\sim}{z}_t - F_t\, \underset{\sim}{\theta}_t \doteq N_m(0,\, U_t(\hat{\underset{\sim}{\theta}}{}_t^C)),$$

$$\hat{\underset{\sim}{z}}{}^P_{t|t-1} - F_t\, \underset{\sim}{\theta}_t \doteq N_m(0,\, F_t\, C_{t|t-1}\, F_t')$$

Par conséquent, sous les hypothèses de notre modèle, comme $\underset{\sim}{z}_t$ et $\hat{\underset{\sim}{z}}{}^P_{t|t-1}$ ne sont pas corrélés étant donné $\underset{\sim}{\theta}_t$, nous obtenons

$$\underset{\sim}{z}_t - \hat{\underset{\sim}{z}}{}^P_{t|t-1} \doteq N_m(0,\, D_{t|t-1}),$$

où $D_{t|t-1}$ est $F_t\, C_{t|t-1}\, F_t' + U_t(\hat{\theta}{}_t^C)$. Le résultat (4.1) en découle immédiatement.

Considérons maintenant la situation des données groupées, dans laquelle l'équation de mesure représente un modèle saturé pour $\underset{\sim}{\mu}_t$, c.-à-d. que le vecteur $\underset{\sim}{\theta}_t$ contient m paramètres. Transversalement, le prédicteur optimal de $\underset{\sim}{y}_t$ est $\underset{\sim}{y}_t$ lui-même. Longitudinalement, on pourrait également montrer qu'en introduisant des paramètres d'écart non aléatoires dans l'équation de transition, $\hat{\underset{\sim}{y}}{}^L_t$ est $\underset{\sim}{y}_t$ lui-même. Par ailleurs, $\hat{\underset{\sim}{z}}{}^P_{t|t-1}$ (ou $\hat{\underset{\sim}{y}}{}^P_{t|t-1}$) est égal à $\underset{\sim}{z}_t$ (ou $\underset{\sim}{y}_t$), ce qui signifie que la SCE est zéro. Ceci est donné par la proposition suivante.

Proposition 4.2 Pour le cas groupé, laissons $\underset{\sim}{\theta}_t$ représenter les paramètres du modèle saturé pour le comportement transversal et $\underset{\sim}{\gamma}_t$ représenter les paramètres d'écart variables dans le temps, inconnus mais non aléatoires, dans la modélisation du comportement longitudinal, c.-à-d. que l'équation de transition est donnée par la formule suivante pour $t \geqslant 2$,

$$\underset{\sim}{\theta}_t = G_t\, \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\gamma}_t + \underset{\sim}{\xi}_t. \tag{4.2}$$

Alors

$$\hat{\underset{\sim}{y}}{}^L_t = \underset{\sim}{y}_t\ , \tag{4.3a}$$

et

$$\hat{\underset{\sim}{y}}{}^P_{t|t-1}(\hat{\underset{\sim}{\gamma}}_t) = \underset{\sim}{y}_t. \tag{4.3b}$$

Pour le prouver, remarquons que la forme réduite (2.14) peut maintenant être modifiée pour contenir les paramètres $\underset{\sim}{\gamma}_2, \ldots, \underset{\sim}{\gamma}_T$, avec $\underset{\sim}{\theta}_T$. Le nombre de paramètres est alors le même que le nombre de z_{it}, c.-à-d. mT. Le résultat (4.3a) en découle naturellement, parce que $\underset{\sim}{y}_t$ est $F_t^{-1}(\underset{\sim}{z}_t)$ pour le modèle saturé. Pour obtenir (4.3b), notons d'abord que pour les $\underset{\sim}{\gamma}_t$ donnés, $\hat{\underset{\sim}{z}}{}^P_{t|t-1}$ est $F_t\, G_t\, \hat{\underset{\sim}{\theta}}{}^L_{t-1} + F_t\, \underset{\sim}{\gamma}_t$. En substituant les estimations de $\underset{\sim}{\gamma}_t$ obtenues par la minimisation de la SCE de (3.5) jusqu'au temps t, on peut voir que $\hat{\underset{\sim}{z}}{}^P_{t|t-1}$ est $\underset{\sim}{z}_t$ et de là (4.3b) en découle.

où $\hat{\boldsymbol{y}}^{P}_{t|t-1}$ est $g^{-1}(g(\boldsymbol{y}_{t-1}) + \hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\beta}}$ dénote la moyenne des différences premières des $g(\boldsymbol{y}_t)$, et k_0 dénote la longueur de $\boldsymbol{\beta}$ dans le modèle (3.6). De même, pour le modèle qui nous intéresse défini par (3.3) et (3.4), nous calculons

$$\mathrm{REQMEW}_1 = [\textstyle\sum_{t=2}^{T}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}^{P}_{t|t-1})'(\boldsymbol{y}_t - \hat{\boldsymbol{y}}^{P}_{t|t-1})/(m(T-1) - k_1)]^{\frac{1}{2}}, \tag{3.8}$$

où k_1 est le nombre de paramètres fixes estimés afin d'appliquer le prédicteur linéaire. Dans le cas des données non groupées, on modifie les dénominateurs dans (3.7) et (3.8) en conséquence. Si REQMEW_1 est plus grand que REQMEW_0, il est nettement inutile de s'attarder sur ce modèle.

3.3.2 Diagnostics post-échantillon

(a) On peut définir des tests prédictifs post- échantillon pour le cas des données groupées. Voir la section suivante pour leur justification asymptotique lorsque n_t et T sont grands. Avec les prédictions de τ pas en avant $\hat{\boldsymbol{z}}^{P}_{T+\tau|T}$, définies ci-dessous à la section 3.4, on procède à un test du chi-carré pour le non-ajustement du modèle en rejetant pour de grandes valeurs de X^2_τ (se rapportant ici à une distribution χ^2_m), où pour $\tau=1, 2, \ldots, T'$,

$$X^2_\tau = (\boldsymbol{z}_{T+\tau} - \hat{\boldsymbol{z}}^{P}_{T+\tau|T})'\, D^{-1}_{T+\tau|T}\, (\boldsymbol{z}_{T+\tau} - \hat{\boldsymbol{z}}^{P}_{T+\tau|T}). \tag{3.9}$$

(b) Les erreurs de validation par recoupement peuvent être calculées pour le modèle simple et pour le modèle examiné, et on peut les étudier pour voir l'ampleur de l'amélioration. Dans le cas des prédictions à τ pas en avant, on peut définir les erreurs de validation par recoupement par la racine de l'écart quadratique moyen des erreurs de prédiction τ pas en avant pour le post- échantillon comme suit

$$\mathrm{REQMEP}_0(\tau) = [\sum_{j=0}^{T'-\tau}(\boldsymbol{y}_{T+\tau+j} - \bar{\boldsymbol{y}}^{P}_{T+\tau+j|T+j})'(\boldsymbol{y}_{T+\tau+j} - \bar{\boldsymbol{y}}^{P}_{T+\tau+j|T+j})/m(T'-\tau+1)]^{\frac{1}{2}} \tag{3.10}$$

où $\bar{\boldsymbol{y}}^{P}_{T+\tau+j|T+j}$ sont les prédictions du modèle simple. De même, on peut définir $\mathrm{REQMEP}_1(\tau)$ pour le modèle qui nous intéresse en utilisant les vecteurs non transformés $\boldsymbol{y}_t$ et leurs prédicteurs.

3.4 Prédiction et lissage

Les prédictions de la période post-échantillon sont nécessaires à des fins de diagnostic et, si le modèle est considéré adéquat, les prédictions pour les observations futures et leurs EQM respectives seraient généralement requises. À cette fin, on laisse simplement de côté les équations de mise à jour du FK, et le MPLN de $\boldsymbol{\theta}$, pour τ périodes en avant, s'obtient d'abord par récursivité en posant

$$\hat{\boldsymbol{\theta}}^{P}_{T+\tau|T} = G_{T+\tau}\, \hat{\boldsymbol{\theta}}^{P}_{T+\tau-1|T}, \tag{3.11}$$

et son EQM par

$$C_{T+\tau|T} = G_{T+\tau}\, C_{T+\tau-1|T}\, G'_{T+\tau} + W_{T+\tau} \tag{3.12}$$

À noter que pour la prédiction à $t>T+T'$, toutes les données jusqu'au point $T+T'$ inclus doivent être utilisées en rajustant le modèle. Or, le prédicteur de $\boldsymbol{z}_{T+\tau}$ est donné par la formule

$$\hat{\boldsymbol{z}}^{P}_{T+\tau|T} = F_{T+\tau}\, \hat{\boldsymbol{\theta}}^{P}_{T+\tau|T}\ , \tag{3.13}$$

et l'EQM correspondante est

$$D_{T+\tau|T} = F_{T+\tau}\, C_{T+\tau|T}\, F'_{T+\tau} + U_{T+\tau}\ . \tag{3.14}$$

On peut donc évaluer $U_{T+\tau}$ en $\hat{\boldsymbol{\theta}}^{P}_{T+\tau|T}$, le prédicteur τ pas en avant de $\theta_{T+\tau}$, chaque fois que $\hat{\boldsymbol{\theta}}^{C}_{T+\tau}$ n'est pas disponible. Pour $\boldsymbol{y}_t$ non transformé, $\hat{\boldsymbol{y}}^{P}_{T+\tau|T}$ est obtenu comme étant $g^{-1}(\hat{\boldsymbol{z}}^{P}_{T+\tau|T})$ et l'EQM pour $\boldsymbol{y} - \hat{\boldsymbol{y}}^{P}_{T+\tau|T}$ est donnée approximativement par $(d\boldsymbol{\mu}_t/d\boldsymbol{\eta}'_t)\ D_{T+\tau|T}\ (d\boldsymbol{\mu}_t/d\boldsymbol{\eta}'_t)'$.

Étape II: Filtrage de Kalman pour l'obtention de $\hat{\underset{\sim}{\theta}}{}^{L}_{T}$

Le MPLN $\hat{\underset{\sim}{\theta}}{}^{L}_{T}$ (approximatif seulement, compte tenu de la linéarisation à l'étape I) de $\underset{\sim}{\theta}_T$ basé sur $\underset{\sim}{z}_1, \dots, \underset{\sim}{z}_T$ peut être calculé de la même façon que $\tilde{\underset{\sim}{\theta}}{}^{L}_{T}$ l'a été à partir du FK donné en (2.16). Les modifications appropriées de (2.16) s'obtiennent en remplaçant $\tilde{\underset{\sim}{\theta}}_t$, V_t et $A_{t|t-1}$ par $\hat{\underset{\sim}{\theta}}_t$, U_t, et $C_{t|t-1}$ respectivement. Le FK est amorcé par $\hat{\underset{\sim}{\theta}}{}^{L}_{1}$ et C_1, où $\hat{\underset{\sim}{\theta}}{}^{L}_{1}$ est l'estimation par les MCP transversale $\hat{\underset{\sim}{\theta}}{}^{C}_{1}$ comme dans (2.7) pour $\underset{\sim}{\theta}_1$ fixé lorsque U_1 est substituée à Σ_1, et C_1 est $(F_1' U_1^{-1} F_1)^{-1}$. Comme on l'a mentionné plus tôt dans l'introduction, le filtre de Kalman généralisé (FKG) proposé par Zehnwirth (1988) pour la variance d'observations dépendantes de l'état et la fonction de lien identité se rattache au FK ci-dessus pour le modèle défini par (3.3) et (3.4) au même sens que Zehnwirth utilise $\overline{U}(= E_\theta\ U_t(\underset{\sim}{\theta}_t))$ et non $U_t(\hat{\underset{\sim}{\theta}}{}^{C}_{t})$ dans la définition du FK. Ceci revient essentiellement à approximer $\overline{U}$ par l'expression à l'intérieur de l'espérance. $\overline{U}$ serait en général impossible à calculer pour une fonction de lien g non linéaire. Toutefois, si U était disponible, elle serait préférable pour des considérations d'optimalité. Nous pouvons également calculer la somme des carrés des erreurs pour le modèle (3.3) analogue à l'expression (2.18), comme un sous-produit du filtrage de Kalman de la façon suivante:

$$SSE = \textstyle\sum_{t=1}^{T}(\underset{\sim}{z}_t - \hat{\underset{\sim}{z}}{}^{P}_{t|t-1})'\ D^{-1}_{t|t-1}\ (\underset{\sim}{z}_t - \hat{\underset{\sim}{z}}{}^{P}_{t|t-1}), \tag{3.5}$$

où $\hat{\underset{\sim}{z}}{}^{P}_{1|0}$ et $D_{1|0}$ sont définis comme $F_1\ \hat{\underset{\sim}{\theta}}{}^{C}_{1}$ et $U_1(\hat{\underset{\sim}{\theta}}{}^{C}_{1})$ respectivement, et $D_{t|t-1}$ comme dans (2.18), où A est remplacée par C et V par U. À partir de la SCE, on peut obtenir une estimation du paramètre de surdispersion σ^2 comme étant SCE/DL, où DL dénote le nombre de degrés de liberté approprié. Après avoir ajusté le modèle, nous examinons maintenant quelques méthodes pour en faire la vérification.

3.3 Diagnostics

Supposons que les données sont disponibles jusqu'à la période $T+T'$. Supposons que les données pour les T premiers points (choisis arbitrairement) servent à l'ajustement du modèle. Nous désignerons des diagnostics basés sur ces points comme "à l'intérieur de l'échantillon" et ceux basés sur les instants $T+1, \dots, T+T'$ comme "post-échantillon". On peut utiliser les moyens suivants pour vérifier l'ajustement du modèle; Harvey (1984), Harvey et Durbin (1986), et Harvey et Fernandes (1989).

3.3.1 Diagnostics à l'intérieur de l'échantillon

(a) Soit r_{it} le résidu de prédiction un pas en avant normalisé correspondant à l'élément i du vecteur $\underset{\sim}{z}_t$ au temps t. Ces résidus, pour chaque i, peuvent être tracés en fonction du temps et en fonction de $\tilde{z}^{P}_{it|t-1}$, et leur comportement aléatoire peut être examiné.

(b) Vérifions si la variance échantillonnale des résidus $\{r_{it}:\ t=2, \dots, T\}$ pour chaque i est proche de un. Une valeur plus grande que un signifie une surdispersion par rapport au modèle en cours d'ajustement (Harvey et Fernandes, 1989).

(c) À la suite de Harvey (1984), on choisit d'abord un modèle simple comme étalon, défini par

$$g(\underset{\sim}{y}_t) = g(\underset{\sim}{y}_{t-1}) + \underset{\sim}{\beta} + \underset{\sim}{\zeta}_t,\ \underset{\sim}{\zeta}_t \sim SL(0, \sigma^2 I), \tag{3.6}$$

où g est la fonction de lien définie par (3.1b) et $\underset{\sim}{\beta}$ est un paramètre d'écart constant. On calcule ensuite la racine de l'erreur quadratique moyenne des erreurs de prédiction (REQME) un pas en avant à l'intérieur de l'échantillon pour le cas des données groupées comme suit

$$REQMEW_0 = [\textstyle\sum_{t=2}^{T}(\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}{}^{P}_{t|t-1})'(\underset{\sim}{y}_t - \tilde{\underset{\sim}{y}}{}^{P}_{t|t-1})/(m(T-1)-k_0)]^{\frac{1}{2}} \tag{3.7}$$

noter que l'estimateur $\hat{W}$, contrairement à $\hat{\hat{W}}$, est toujours défini non négatif, ce qui est naturellement souhaitable dans la pratique. Toutefois, à moins que T ne soit grand, $\hat{W}$ ne sera pas convergent pour W. On verra plus tard à la section 4 que, dans des conditions assez peu rigoureuses, ce type de mauvaise spécification de W quand T n'est pas grand ne touche pas la convergence des estimations des paramètres du prédicteur du MLGEÉ .

3. LA MÉTHODE PROPOSÉE - Le MLGEÉ

3.1 Définition Le modèle linéaire généralisé à espace d'états (MLGEÉ) peut être défini par les deux équations suivantes.

(i) Comportement transversal: Pour chaque $t=1, \ldots, T$,

$$\underline{y}_t = \underline{\mu}_t + \underline{\varepsilon}_t, \tag{3.1a}$$

$$\underline{\eta}_t \equiv g(\underline{\mu}_t) = F_t \, \underline{\theta}_t, \tag{3.1b}$$

où $\underline{\varepsilon}_t | \underline{\mu}_t \sim SL(0, V_t(\underline{\mu}_t))$, $Cov(\underline{\varepsilon}_t, \underline{\varepsilon}_s | \underline{\mu}_t, \underline{\mu}_s) = 0$ for $t \neq s$, et g est une fonction de lien monotone et différentiable .

(ii) Comportement longitudinal: Pour $t=2, \ldots, T$,

$$\underline{\theta}_t = G_t \, \underline{\theta}_{t-1} + \underline{\xi}_t, \tag{3.2}$$

où $\underline{\xi}_t \sim SL(0, W_t)$ avec les conditions habituelles données plus tôt par (2.14).

Les deux principales différences entre cette formulation et celle du MLEÉ donnée par (2.11) et (2.13) consistent en ce que la matrice des variances et covariances V_t dépend du vecteur moyenne $\underline{\mu}_t$, et par conséquent du vecteur d'états $\underline{\theta}_t$, et que la fonction de lien n'est pas nécessairement l'identité. En ajustant le MLGEÉ aux données des séries chronologiques, on supposera en général que n_t et T sont grands. Le choix des matrices de plan F_t peut être justifié par des analyses transversales et celui des G_t par l'analyse de la structure temporelle dans la série des estimations transversales $\{\hat{\underline{\theta}}_t^C\}$. La matrice des variances et covariances W_t, si elle n'est pas connue a priori, peut être estimée par $\hat{W}$ sous l'hypothèse de l'invariance dans le temps, telle que décrite à la section 2.4. On notera également que la formulation ci-dessus peut être évidemment prolongée afin de prendre en compte le paramètre de surdispersion σ^2, comme cela était le cas avec le MLEÉ examiné plus haut à la partie 2.3. Pour ajuster le MLGEÉ, nous proposons l'algorithme suivant pour l'estimation des paramètres du modèle.

3.2 Algorithme d'estimation - MCPIF

L'algorithme des moindres carrés pondérés par itérations et filtrage (MCPIF) pour l'estimation (ou la prédiction) de $\underline{\theta}_T$ comprend deux étapes, chacune nécessitant une série de pas itératifs.

Étape I: Linéarisation pour la formulation de l'espace d'états

Transformons d'abord $\underline{y}_t$ en $\underline{z}_t$ pour chaque $t=1, \ldots, T$ comme en (2.5). Maintenant, pour n_t grand, on peut définir un contexte de travail de MLEÉ approximatif pour les séries $\{\underline{z}_t\}$ par

$$\underline{z}_t = F_t \, \underline{\theta}_t + \underline{\delta}_t, \tag{3.3a}$$

$$\underline{\theta}_t = G_t \, \underline{\theta}_{t-1} + \underline{\xi}_t, \tag{3.3b}$$

où

$$\underline{\delta}_t \doteq SL(0, U_t(\hat{\underline{\theta}}_t^C)), \; \underline{\xi}_t \sim SL(0, W_t), \tag{3.4a}$$

$$U_t(\hat{\underline{\theta}}_t^C) = (d\underline{\eta}_t / d\underline{\mu}_t') \, V_t(\underline{\mu}_t)(d\underline{\eta}_t / d\underline{\mu}_t')' \Big|_{\underline{\theta}_t = \hat{\underline{\theta}}_t^C} . \tag{3.4b}$$

Les vecteurs d'erreurs $\underline{\delta}_t$, $\underline{\xi}_t$ répondent aux conditions habituelles données plus tôt pour la définition du MLEÉ à la partie 2.3.

Dans la section suivante, nous proposons les modèles linéaires généralisés à espace d'états, MLGEÉ, comme un prolongement des MLEÉ. À noter que dans le prolongement en MLG du ML, on a linéarisé le modèle en transformant de $\underset{\sim}{y}_t$ à $\underset{\sim}{z}_t$ par la méthode des MCPI. Il est donc naturel de définir le MLGEÉ en appliquant le MLEÉ à la série transformée $\{\underset{\sim}{z}_t\}$, c.-à-d. que l'algorithme des MCPF est administré sur $\{\underset{\sim}{z}_t\}$. En d'autres termes, il faut procéder par filtrage et itérations afin d'obtenir les estimations par les MCP dans le MLGEÉ. Ceci mène à l'algorithme des MCPIF pour la méthode proposée. On peut constater que cet algorithme est assez proche de l'algorithme des MCFPI (moindres carrés filtrés et pondérés de façon itérative) de Zeger (1988), qui avait été introduit dans un but différent et qui n'utilise pas le filtre de Kalman récurrent. En utilisant les MCPIF, nous devons d'abord spécifier les matrices des variances et covariances des erreurs $V_t(\underset{\sim}{\mu}_t(\theta_t))$ et W_t. Pour n_t grand, la matrice V_t peut être assez bien approximée par $V_t(\underset{\sim}{\mu}_t(\hat{\underset{\sim}{\theta}}{}_t^C))$ où $\hat{\underset{\sim}{\theta}}{}_t^C$ est une estimation convergente de $\underset{\sim}{\theta}_t$ semblable à celle donnée dans (2.7). En ce qui concerne W_t, si nous pouvons supposer qu'il est invariant dans le temps, c.-à-d. que $W_t = W$, alors pour T grand il est possible de construire un estimateur convergent en utilisant une méthode parallèle à celle utilisée dans les modèles de régression à coefficients aléatoires (RCA) de Swamy (1970) qui est décrite dans la sous-section suivante.

2.4 Spécification de la matrice des variances et covariances W_t sous l'hypothèse d'invariance dans le temps

On peut définir un estimateur convergent $\hat{W}$ sous l'hypothèse $W_t = W$ lorsque n_t et T sont grands. Dans les modèles de régression avec coefficients aléatoires proposés pour l'économétrie des données transversales, Swamy (1970) a utilisé les estimations de la régression par les moindres carrés $\hat{\beta}_i$ provenant de plusieurs groupes (ou grappes) afin d'estimer la variance de la composante de régression aléatoire β_i; voir aussi Pfeffermann et Nathan (1981). Bien que le problème de la prédiction avec les séries chronologiques soit tout à fait différent de celui de l'estimation du β sous-jacent (ou d'une fonction des β_i), les estimations transversales convergentes $\{\hat{\theta}_t^C,\ t=1, \dots T\}$ peuvent être utilisées de la même façon pour estimer W. Dans la méthode de Swamy (1970), on obtient une estimation de la variance corrigée pour le biais. Dans le contexte de notre travail, la propriété d'abscence de biais correspondrait à l'absence de biais asymptotique pour n_t grand. L'estimateur $\hat{W}$ peut être défini comme suit.

Pour $t \geqslant 2$, soit

$$\hat{\beta}_t = \hat{\underset{\sim}{\theta}}{}_t^C - G_t\, \hat{\underset{\sim}{\theta}}{}_{t-1}^C \tag{2.20a}$$

$$R_1 = (T-1)^{-1} \textstyle\sum_{t=2}^T E[(\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)\ (\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)'] \tag{2.20b}$$

$$R_2 = (T-1)^{-1} \textstyle\sum_{t=2}^T E\ [\underset{\sim}{\beta}_t\ (\hat{\underset{\sim}{\beta}}_t - \underset{\sim}{\beta}_t)'], \tag{2.20c}$$

et définissons deux estimateurs $\hat{W}$ et $\hat{\hat{W}}$ donnés par

$$\hat{W} = (T-1)^{-1} \textstyle\sum_{t=2}^T \hat{\underset{\sim}{\beta}}_t \hat{\underset{\sim}{\beta}}{}_t', \quad \hat{\hat{W}} = \hat{W} - R_1 - R_2 - R_2'. \tag{2.20d}$$

Nous avons

$$E(\hat{W}) = W + R_1 + R_2 + R_2', \quad E(\hat{\hat{W}}) = W \tag{2.21}$$

Le biais de $\hat{W}$ est donc donné par $R_1+R_2+R_2'$. À noter que le terme R_2 n'est pas nul parce que $E(\hat{\underset{\sim}{\beta}}_t | \underset{\sim}{\theta}_t,\ \underset{\sim}{\theta}_{t-1})$ n'est pas en général égale à $\underset{\sim}{\beta}_t$. L'estimateur corrigé pour le biais $\hat{\hat{W}}$ de (2.20d) avec des estimations appropriées des R_i est analogue à l'estimateur de la variance de Swamy (1970). Toutefois, le terme de biais $R_1+R_2+R_2'$ serait négligeable pour n_t grand, lorsque la moyenne et la covariance de $\hat{\underset{\sim}{\beta}}_t$ conditionnelles sur $(\underset{\sim}{\theta}_t,\ \underset{\sim}{\theta}_{t-1})$ coïncident en limite avec celles de la distribution asymptotique. Dans cet article on posera par hypothèse les conditions de régularité nécessaires pour que ceci soit vrai, et par conséquent $\hat{W}$ sera (approximativement) sans biais. Donc, pour n_t grand, nous pouvons éliminer la correction pour le biais et utiliser simplement $\hat{W}$ pour estimer W. Il est à méthode des MCP grâce à un FK avec une distribution a priori stable pour le vecteur d'états initial, parce que les estimations ainsi obtenues par récurrence sont MPLN (meilleurs prédicteurs linéaires non biaisés) ou ELNMMQ (estimations linéaires non biaisées minimisant la moyenne quadratique), voir Harvey (1981, p. 105 et Zehnwirth (1988)). Une modification appropriée de la distribution du vecteur d'états initial sera nécessaire, puisqu'à l'état initial certains des éléments sont stationnaires; voir Harvey et Peters (1984). L'algorithme de récurrence pour le FK qui donne le MPLN $\tilde{\underline{\theta}}^P_{t|t-1}$ de $\underline{\theta}_t$ étant donné $\underline{y}_1, \ldots, \underline{y}_{t-1}$, et le prédicteur mis à jour $\tilde{\underline{\theta}}^L_t$, étant donné $\underline{y}_1, \ldots, \underline{y}_t$ pour chaque $t \geq 2$, est donné par la formule

$$\tilde{\underline{\theta}}^P_{t|t-1} = G_t \, \tilde{\underline{\theta}}^L_{t-1}, \tag{2.16a}$$

$$\tilde{\underline{\theta}}^L_t = \tilde{\underline{\theta}}^P_{t|t-1} + K_t(\underline{y}_t - F_t \, \tilde{\underline{\theta}}^P_{t|t-1}), \tag{2.16b}$$

$$K_t = A_{t|t-1} \, F_t' \, (F_t \, A_{t|t-1} F_t' + V_t)^{-1}, \tag{2.16c}$$

$$A_{t|t-1} = G_t A_{t-1} \, G_t' + W_t, \tag{2.16d}$$

$$A_t = (I - K_t F_t) \, A_{t|t-1}, \tag{2.16e}$$

où $A_{t|t-1}$ est la matrice des variances et covariances non conditionnelles des erreurs de $\tilde{\underline{\theta}}^P_{t|t-1}$, c.-à-d. son EQM (erreur quadratique moyenne), et $A_t = A_{t|t}$, c.-à-d. l'EQM de $\tilde{\underline{\theta}}^L_t$. On peut voir que les valeurs de $\tilde{\underline{\theta}}^L_1$ et A_1 pour commencer avec le FK (2.16) sont $\tilde{\underline{\theta}}^C_1$ et $(F_1' \, V_1^{-1} \, F_1^{-1})$ respectivement, où $\tilde{\underline{\theta}}^C_1$ est l'estimation par les MCP transversale (2.2) pour $\underline{\theta}_1$ fixé, quand Σ_1 est remplacée par V_1. La matrice K_t est le gain de Kalman au temps t. L'algorithme ci-dessus donne également par récursivité les distributions au sens large de $(\tilde{\underline{\theta}}^L_t - \underline{\theta}_t)$, $t=1, \ldots T$ dans le calcul de $\tilde{\underline{\theta}}^L_T$. Cela pour $t=1, \ldots T$.

$$\tilde{\underline{\theta}}^L_t - \underline{\theta}_t \sim SL(0, A_t). \tag{2.17}$$

À titre d'analogie avec la méthode des MCPI utilisée dans le calcul de $\tilde{\underline{\theta}}^C_t$ pour les MLG, la méthode ci-dessus du calcul de $\tilde{\underline{\theta}}^L_t$ pour les MLEÉ par les MCP grâce au filtre de Kalman sera désignée dans cet article comme la méthode des MCPF, afin de faire ressortir sa relation avec la méthode habituelle des MCP pour les ML.

En plus de fournir divers MPLN, le filtre de Kalman donne également une méthode simple de calcul de la somme des carrés des erreurs pour le modèle (2.14) ou (2.11) et (2.13), grâce aux résidus de prédiction un pas en avant $\underline{y}_t - \tilde{\underline{y}}^P_{t|t-1}$ et à leurs EQM. Il s'ensuit du résultat d'équivalence (B.2) prouvé par Harvey et Peters (1984) que pour tout $t'=\tau$,

$$(\underline{y} - F^{*'}_\tau \, \tilde{\underline{\theta}}^L_\tau)' \, \Omega^{-1}_\tau \, (\underline{y} - F^{*'}_\tau \, \tilde{\underline{\theta}}^L_\tau) = \sum_{t=1}^T \, (\underline{y}_t - \tilde{\underline{y}}^P_{t|t-1})' B^{-1}_{t|t-1} \, (\underline{y}_t - \tilde{\underline{y}}^P_{t|t-1}), \tag{2.18}$$

où pour $t \geq 2$,

$$\tilde{\underline{y}}^P_{t|t-1} = F_t \, \tilde{\underline{\theta}}^P_{t|t-1}, \; B_{t|t-1} = F_t \, A_{t|t-1} \, F_t' + V_t \tag{2.19}$$

et $\tilde{\underline{y}}^P_{1|0}$ et $B_{1|0}$ sont posés égaux à $F_1 \, \tilde{\underline{\theta}}^C_1$ et V_1 respectivement. Le résultat ci-dessus est analogue à l'équivalence de la SCE provenant des moindres carrés ordinaires pour les ML et la somme des carrés des résidus de prédiction un pas en avant obtenus par la méthode des moindres carrés récursive.

Enfin, on peut voir facilement à partir de (2.16) que les matrices des variances et covariances du modèle V_t et W_t ne sont spécifiées que jusqu'à un paramètre de surdispersion multiplicatif σ^2, les estimations $\tilde{\underline{\theta}}^P_1$ et $\tilde{\underline{\theta}}^L_1$ ne changent pas, sauf pour la correction multiplicative de leur EQM par un facteur de σ^2.

l'a déjà dit, est soit un n_t-vecteur pour les données non groupées, soit un m-vecteur dans le cas des données groupées. On utilise deux équations pour modéliser avec les MLEÉ, voir p. ex. Zehnwirth (1988). D'abord, pour le comportement transversal, on définit l'<u>équation de mesure</u> comme:

$$\underset{\sim}{y}_t = F_t \, \underset{\sim}{\theta}_t + \underset{\sim}{\varepsilon}_t, \tag{2.11}$$

où F_t est une matrice connue de variables auxiliaires, $\underset{\sim}{\theta}_t$ est un r-vecteur de paramètres aléatoires que l'on appelle vecteur d'états, et la distribution des erreurs aléatoires $\underset{\sim}{\varepsilon}_t$ jusqu'aux moments du deuxième ordre est

$$\underset{\sim}{\varepsilon}_t | \underset{\sim}{\theta}_t \sim SL(0, V_t), \; Cov(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_s | \underset{\sim}{\theta}_t \, \underset{\sim}{\theta}_s) = 0 \text{ pour } t \neq s. \tag{2.12}$$

La matrice des variances et covariances V_t ne dépend pas de $\underset{\sim}{\theta}_t$ et elle est supposée connue pour tout t. Ensuite, dans le cas du comportement longitudinal, on définit l'<u>équation de transition</u> comme

$$\underset{\sim}{\theta}_t = G_t \, \underset{\sim}{\theta}_{t-1} + \underset{\sim}{\xi}_t, \tag{2.13a}$$

où G_t est une matrice de transition $r \times r$ connue, et les erreurs $\underset{\sim}{\xi}_t$ sont spécifiées par

$$\underset{\sim}{\xi}_t \sim SL(0, W_t), \; Cov(\underset{\sim}{\xi}_t, \underset{\sim}{\xi}_s) = 0, \; s \neq t, \text{ et}$$

$$Cov(\underset{\sim}{\xi}_t, \underset{\sim}{\varepsilon}_s | \underset{\sim}{\theta}_s) = 0 \text{ pour tout } s,t; \; Cov(\underset{\sim}{\xi}_t, \theta_s) = 0 \text{ pour } t > s. \tag{2.13b}$$

On suppose également que la matrice des variances et covariances W_t est connue. On remarquera que l'hypothèse du type Markov dans l'équation de transition (2.13) touche l'estimation récurrente et n'est pas nécessaire pour des considérations d'optimalité.

Le modèle défini par (2.11) et (2.13) est complètement spécifié, sauf pour la distribution du vecteur d'états initial $\underset{\sim}{\theta}_0$. Ici, nous ne considérerons pas les méthodes d'initialisation habituelles décrites dans Harvey (1981, ch. 4) et Harvey et Peters (1984), qui sont alors suivies de l'estimation optimale des paramètres $\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2, \ldots, \underset{\sim}{\theta}_T$ successivement par le filtre de Kalman (FK). À la place, nous considérerons d'abord une forme réduite de (2.11) et de (2.13) en une équation unique ne contenant qu'un vecteur de paramètres $\underset{\sim}{\theta}_T$ et ensuite une méthode appropriée pour estimer $\underset{\sim}{\theta}_T$, qui sera nécessaire pour la prédiction de $\underset{\sim}{y}_t$ pour $t > T$. Cette approche sera utile pour mettre en relation le MLEÉ avec le ML et le MLG décrits plus haut.

Conditionnellement sous $\underset{\sim}{\theta}_T$, on peut écrire les modèles (2.11) et (2.13) comme un ML pour $\underset{\sim}{y} = (\underset{\sim}{y}_1', \ldots, \underset{\sim}{y}_T')'$ comme dans Harvey et Peters (1984). En écrivant $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_{T-1}$ en termes de $\underset{\sim}{\theta}_T$ et des $\underset{\sim}{\xi}_t$, nous obtenons

$$\underset{\sim}{y} = F_T^* \, \underset{\sim}{\theta}_T + \underset{\sim}{\xi}_T^*, \tag{2.14}$$

où F_T^* est une matrice $T^* \times r$ connue de valeurs fixes (l'ordre T^* sera mT dans le cas des données groupées et $n_1 + \ldots + n_T$ dans le cas des données non groupées), et $\underset{\sim}{\varepsilon}_T^*$ est un nouveau vecteur d'erreurs $T^* \times 1$ de moyenne zéro et de matrice des variances et covariances Ω_T. La matrice Ω_T peut être complètement spécifiée en termes des matrices connues V_t, W_t, F_t et G_t. À noter que le modèle (2.14) aurait pu être écrit conditionnellement sous $\underset{\sim}{\theta}_\tau$ à tout point dans le temps $t = \tau$. Ainsi, $\underset{\sim}{\theta}_T$ peut être estimé de façon optimale en utilisant les MCP comme dans le ML par l'expression

$$\underset{\sim}{\overset{\sim}{\theta}}{}_T^L = (F_T^{*\prime} \, \Omega_T^{-1} \, F_T^*)^{-1} \, F_T^{*\prime} \, \Omega_T^{-1} \, \underset{\sim}{y}, \tag{2.15}$$

où L représente les données longitudinales utilisées dans l'estimation.

L'expression ci-dessus comporte l'inversion de Ω_T, qui serait en général de grande dimension, ce qui pourrait par conséquent se traduire par des difficultés de calcul. On peut cependant facilement évaluer l'estimation par la

On reprend le processus ci-dessus jusqu'à la convergence. En désignant la solution convergente $\tilde{\underset{\sim}{\theta}}_t^C$, $\underset{\sim}{z}_t$ la variable correspondante de l'expression (2.5) et Γ_t la matrice correspondante de (2.6), nous avons, à mesure que $n_t \to \infty$,

$$\tilde{\underset{\sim}{\theta}}_t^C \;\dot{\sim}\; N_r\,(\underset{\sim}{\theta}_t\,,\,(D_t' \,\Sigma_t^{-1}\, D_t)^{-1}), \tag{2.7}$$

et

$$\underset{\sim}{z}_t \;\dot{\sim}\; SL\,(\underset{\sim}{\eta}_t\,,\,\Gamma_t), \tag{2.8}$$

où (2.7) est valide sous une application appropriée du TLC. Remarquons que la longueur de $\underset{\sim}{z}_t$ augmente avec n_t dans le cas non groupé, où la distribution asymptotique dans (2.8) doit être interprétée en termes de toutes les marginales de dimension finie de $\underset{\sim}{z}_t$. Les équations (2.7) et (2.8) ci- dessus sont des MLG analogues à (2.3) et à (2.1) respectivement en ce sens que

$$(D_t' \,\Sigma_t^{-1}\, D_t)^{-1} = (F_t' \,\Gamma_t^{-1}\, F_t)^{-1}, \tag{2.9}$$

parce que

$$D_t = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\theta}_t') = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}_t')(d\underset{\sim}{\eta}_t'/d\underset{\sim}{\theta}_t') = (d\underset{\sim}{\mu}_t/d\underset{\sim}{\eta}_t')\,F_t, \tag{2.10a}$$

et

$$\Gamma_t^{-1} = (d\underset{\sim}{\mu}_t \,/\, d\underset{\sim}{\eta}_t')' \;\Sigma_t^{-1}\; (d\underset{\sim}{\mu}_t \,/\, d\underset{\sim}{\eta}_t'). \tag{2.10b}$$

De plus, l'estimation $\tilde{\underset{\sim}{\theta}}_t^C$ ne change pas en présence du paramètre de surdispersion σ^2, c.-à-d. lorsque $Cov(\underset{\sim}{y}_t)$ est $\sigma^2\ \Sigma_t\ (\underset{\sim}{\mu}_t)$. La variance asymptotique de $\tilde{\underset{\sim}{\theta}}_t^C$ dans (2.7), par contre, varie par un facteur multiplicatif de σ^2.

Chaque fois que l'observation $\underset{\sim}{y}_t$ donne un estimateur convergent de $\underset{\sim}{\mu}_t$ (ce qui serait le cas, par exemple, si les n_t observations étaient groupées en m blocs), on peut alors utiliser un autre estimateur un pas en avant ("one-step estimator"), $\underset{\sim}{\theta}_t^\star$, suivant la méthodologie GSK de Grizzle, Starmer et Koch (1969). En d'autres termes, l'itération s'arrête après un cycle seulement et ne reprend pas jusqu'à la convergence. On peut montrer que l'estimateur $\underset{\sim}{\theta}_t^\star$ est asymptotiquement équivalent à $\tilde{\underset{\sim}{\theta}}_t^C$. Toutefois, l'estimateur $\tilde{\underset{\sim}{\theta}}_t^C$ serait preférable pour des considérations d'échantillon fini, puisque $\Sigma_t\ (\underset{\sim}{\mu}_t^{(0)})$ peut être instable en raison de la présence de cellules avec un nombre possiblement petit d'observations. Il est intéressant de constater que lorsque H_{2t} est un modèle saturé du cas des données groupées, c.-à-d. lorsque $r = m$, alors les deux estimateurs $\tilde{\underset{\sim}{\theta}}_t^C$ et $\underset{\sim}{\theta}_t^\star$ coïncident l'un avec l'autre et sont égaux à $\underset{\sim}{\theta}_t^{(0)}$ ou à $F_t^{-1}g(\underset{\sim}{y}_t)$.

2.3 La méthode des MCPF (moindres carrés pondéres filtrés) de la théorie des MLEÉ

Nous considérons maintenant la généralisation de la théorie des ML aux MLEÉ (modèles linéaires à espace d'états) afin de prendre en compte la dépendance sérielle. Dans les MLEÉ (voir p. ex. Harvey 1981, chapitre 4 et Harvey, 1984), la dépendance sérielle est introduite au moyen de paramètres $\underset{\sim}{\theta}_t$ variant aléatoirement, $t=1, \ldots, T$, qu sont reliés par des modèles à espace d'états. Dans le problème examiné danscet article, où n_t et T sont tous les deux supposés grands, il semble naturel et pratique d'essayer de modéliser $\underset{\sim}{y}_t$ simultanément pour le comportement transversal, et ensuite de modéliser les paramètres $\underset{\sim}{\theta}_t$ sous-jacents temporellement pour le comportement longitudinal comme dans la modélisation à espace d'états, c.-à-d. que le modèle est spécifié par deux équations. Dans cette sous-section, on résume les méthodes d'estimation pour les MLEÉ. Contrairement aux MLG, la variance ne peut pas varier avec la moyenne, et on n'utilise que la fonction de lien identité. Toutefois, la méthode proposée dans la section suivante généralise les MLEÉ de la même façon que les MLG prolongent les ML afin de donner une méthode appropriée pour le problème décrit plus tôt dans l'introduction.

Contrairement aux deux sous-sections précédentes, nous considérons les données transversales et longitudinales ensemble, c.-à-d. la série chronologique des vecteurs d'observations $\underset{\sim}{y}_t$, $k=1, \ldots T$. Le vecteur $\underset{\sim}{y}_t$, comme on

2.1 La méthode des moindres carrés pondérés (MCP) de la théorie des ML

Pour la coupe transversale au temps t, considérons le prédicteur linéaire ou le modèle H_{1t}: $\underline{\mu}_t = F_t \, \underline{\theta}_t$, où $\underline{\theta}_t$ est un r-vecteur d'effets fixes ($r \leq m$ dans le cas des données groupées) et F_t est une matrice de variables auxiliaires (ou concomitantes) connue. Supposons par ailleurs que Σ_t est constante, c.-à-d. qu'elle ne varie pas avec $\underline{\mu}_t$ et qu'elle soit approximativement connue pour n_t grand. L'estimation optimale $\underline{\tilde{\theta}}_t$ de $\underline{\theta}_t$ au sens de Gauss-Markov donnée par la méthode des MCP est la solution de

$$F_t' \, \Sigma_t^{-1} \, (\underline{y}_t - \underline{\mu}_t) = 0, \tag{2.2a}$$

ce qui implique que

$$\underline{\tilde{\theta}}_t = (F_t' \, \Sigma_t^{-1} \, F_t)^{-1} \, F_t' \, \Sigma_t^{-1} \, \underline{\mu}_t. \tag{2.2b}$$

La distribution asymptotique de $\underline{\tilde{\theta}}_t$ jusqu'aux termes d'ordre n_t^{-1} à mesure que $n_t \to \infty$ est donnée, sous l'application appropriée du TLC (théorème de la limite centrale), par

$$\underline{\tilde{\theta}}_t \doteq N_r \, (\underline{\theta}_t, \, (F_t' \, \Sigma_t^{-1} \, F_t)^{-1}). \tag{2.3}$$

Si $\mathrm{Cov}(\underline{y}_t)$ est connue jusqu'à un multiple constant σ^2 de Σ_t, l'estimateur optimal (2.2b) ne change pas, mais la covariance (2.3) est multipliée par le facteur σ^2, qui est le paramètre de surdispersion ("overdispersion parameter").

2.2 La méthode des MCPI (moindres carrés pondérés itérative) de la théorie des MLG

Nous considérons ensuite la généralisation de la théorie des ML à la théorie des MLG, dans laquelle Σ_t peut varier avec $\underline{\mu}_t$, mais ce de façon connue, et où $\underline{\mu}_t$ peut être une fonction non linéaire monotone et différentiable de $\underline{\theta}_t$, que l'on désigne fonction de lien inverse ("inverse-link function"). La forme de la relation variance-moyenne provient d'une distribution de famille exponentielle, ce qui est analogue à l'hypothèse de la variance constante propre aux distributions normales. On peut estimer $\underline{\theta}_t$ pour les MLG de la façon suivante.

Ici, comme plus haut, posons que t est fixe. Le prédicteur linéaire, après transformation par la fonction de lien, est spécifié par le modèle H_{2t}: $g(\underline{\mu}_t) = F_t \, \underline{\theta}_t$, où $\underline{\theta}_t$ est encore un r-vecteur d'effets fixes, F_t est une matrice connue de variables auxiliaires et g est la fonction de lien. De plus, on suppose que la matrice des variances et covariances $\Sigma_t \, (\underline{\mu}_t)$ de $\underline{y}_t$ est une fonction connue de $\underline{\mu}_t$. Une estimation $\underline{\tilde{\theta}}_t^C$ asymptotiquement optimale (au sens étendu de Gauss-Markov, McCullagh, 1983) de $\underline{\theta}_t$ est donnée par la solution de l'équation de score de quasi-vraisemblance suivante (McCullagh et Nelder, 1989, ch. 9):

$$D_t' \, \Sigma_t^{-1} \, (\underline{Y}_t - \underline{\mu}_t) = 0, \tag{2.4}$$

où Σ_t étant une fonction de $\underline{\mu}_t$ dépend de $\underline{\theta}_t$, et D_t est la matrice $n_t x r$ $(d\underline{\mu}_t / d\underline{\theta}_t')$. Dans le cas groupé, D_t serait une matrice $m \times r$. On peut résoudre l'équation (2.4) par la méthode des MCPI à partir de la procédure de Newton-Raphson. À cette fin, on définit d'abord une variable dépendante ajustée $\underline{z}_t^{(i)}$ pour chaque itération i, $i=1, 2 \ldots$ comme suit.

$$\underline{z}_t^{(i)} = \underline{\eta}_t^{(i-1)} + (d\underline{\eta}_t / d\underline{\mu}_t') \; (\underline{y}_t - \underline{\mu}_t) \; \Big|_{\underline{\theta}_t \, = \, \underline{\theta}_t^{(i-1)}}, \tag{2.5}$$

où $\underline{\eta}_t = g(\underline{\mu}_t)$, et $\underline{\mu}_t^{(0)}$ est posé égal à $\underline{y}_t$. Certaines modifications de $\underline{y}_t$ peuvent être nécessaires si $\underline{\eta}_t^{(0)}$ n'est pas bien défini. Pour chaque itération i, on obtient une estimation $\underline{\theta}_t^{(i)}$ par la méthode des MCP en utilisant le modèle de travail $E \; (\underline{z}_t^{(i)}) = \underline{\eta}_t^{(i)} = F_t \underline{\theta}_t^{(i)}$ et en plus la covariance de travail de $\underline{z}_t^{(i)}$ donnée par

$$\Gamma_t^{(i)} = (d\underline{\eta}_t / d\underline{\mu}_t') \; \Sigma_t (d\underline{\eta}_t / d\underline{\mu}_t')' \; \Big|_{\underline{\theta}_t \, = \, \underline{\theta}_t^{(i-1)}} \; . \tag{2.6}$$

à espace d'états non normale dans laquelle les densités non normales à chaque pas du filtre de Kalman font l'objet d'une évaluation numérique, à Zeger (1988), qui utilise l'estimation d'une équation où l'autocorrélation est introduite par un processus de mélange aléatoire, et récemment à Harvey et Fernandes (1989), sous la forme d'une modélisation à espace d'états non bayesienne, quoique des distributions a priori conjuguées servent à spécifier les équations de transition.

Les méthodes de séries chronologiques pour les données non normales ont été mises au point pour des données univariées ou unidimensionnelles dans le cas qualitatif. Bien qu'il soit possible d'étendre ces méthodes aux données multivariées (ou multidimensionnelles), les calculs nécessaires semblent être assez complexes. Nous proposons ici une solution plus simple dans laquelle le nombre (n_t) d'observations à chaque point dans le temps et le nombre total (T) de points dans le temps sont raisonnablement grands. Même si T n'est pas grand, les estimations des paramètres des modèles restent cohérentes (quand n_t est grand) sous des conditions assez modérées. On appelle le modèle proposé modèle linéaire généralisé à espace d'états (MLGEÉ), dans lequel la méthode du filtre de Kalman est modifiée de façon à s'ajuster à une modélisation non normale et non linéaire. Le filtre de Kalman modifié utilisé dans le MLGEÉ se rattache au filtre généralisé de Kalman de Zehnwirth (1988), lorsque la fonction de lien ("link function") du MLGEÉ est l'identité. Si n_t est grand, on peut "linéariser" le problème afin d'employer les méthodes de modèles linéaires à espace d'états bien connues. Cet élément est semblable à l'idée de la transformation de Smith et Brunsdon (1989). Par ailleurs, n_t grand va donner des estimations de paramètres transversales cohérentes, qui peuvent servir à spécifier la dépendance sérielle entre les observations grâce à l'équation de transition dans la modélisation à espace d'états. Cet aspect se rattache dans une certaine mesure à la méthode utilisée dans Stram, Wei et Ware (1988).

La partie 2 contient d'abord quelques remarques préliminaires touchant notamment la notation et la motivation. On voit que la formulation MLGEÉ se présente presque naturellement pour notre problème. À la partie 3, on définit la méthode MLGEÉ proposée sous un ensemble général d'hypothèses semblables à celles du MLG. La partie 4 contient quelques résultats théoriques, que suit un exemple numérique de projection d'une série de données sur la mortalité due au cancer à la partie 5. La partie 6, enfin, contient une discussion et suggère des directions pour les travaux futurs.

2. REMARQUES PRÉLIMINAIRES

Soit $\underline{y}_t$ le n_t-vecteur des observations au temps t, t = 1,2 ... T. Si les n_t observations sont groupées ou regroupées selon certaines variables en m domaines ou groupes d'intérêt, $\underline{y}_t$ serait également utilisé pour dénoter le m-vecteur des estimations, c'est-à-dire des comptes, des proportions ou des moyennes. On supposera que les éléments du vecteur $\underline{y}_t$ pour le cas non groupé sont indépendants. Toutefois, dans le cas où les observations sont groupées, ces éléments peuvent être dépendants. Dans ce qui suit, on suppose que n_t et T sont grands. Les symboles "~" et "$\dot{\sim}$" serviront à dénoter les expressions "distribué comme" et " asymptotiquement distribué comme" respectivement. Compte tenu du cadre général que nous voulons adopter, nous ne travaillerons qu'avec les hypothèses du deuxième moment, c'est-à-dire les distributions ne seront spécifiées qu'au sens large (SL) seulement. Supposons

$$\underline{y}_t \sim SL\ (\underline{\mu}_t,\ \Sigma_t) \tag{2.1a}$$

et

$$\underline{y} = (\underline{y}_1',\ \ldots,\ \underline{y}_T') \sim SL\ (\underline{\mu},\ \Omega), \tag{2.1b}$$

où on suppose que Σ_t est non singulière et peut varier avec $\underline{\mu}_t$. Par ailleurs, Ω ne sera pas en général une matrice diagonale en blocs en raison de la dépendance sérielle dans les séries chronologiques des vecteurs d'observations $\underline{y}_t$, t=1, ... T. À noter que si les n_t observations ne sont pas groupées, alors Σ_t serait une matrice diagonale en raison de l'hypothèse d'indépendance des observations. Le problème qui nous intéresse est de prédire $\underline{y}_t$ pour t>T. À cette fin, il nous faut un modèle convenable pour $\underline{\mu}$ comme fonction d'un ensemble parcimonieux de paramètres $\underline{\theta}$ de sorte que les $\underline{\mu}$ soient aussi proches que possible des $\underline{y}$.

Nous définissons d'abord certaines notations et certains termes provenant des modèles linéaires (ML), des modèles linéaires généralisés (MLG), des modèles linéaires à espace d'états (MLEÉ) et des modèles de régression à coefficients aléatoires (MRCA). Ces notations nous seront utiles pour justifier la méthode proposée que l'on décrit dans la partie suivante.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ANALYSE DE SÉRIES CHRONOLOGIQUES QUALITATIVES EN TABLEAUX CROISÉS

A.C. Singh et G.R. Roberts[1]

RÉSUMÉ

On propose un cadre avec paramètres pour la définition de modèles linéaires généralisés destinés aux données des séries chronologiques. La structure diachronique des paramètres transversaux est déterminée en fonction de modèles à espace d'états. À cette fin, on utilise des estimations transversales convergentes des paramètres du modèle. Une variante du filtre de Kalman, dans laquelle le vecteur des observations est transformé judicieusement, sert à définir les équations récursives pour la prédiction et la mise à jour. L'application de la méthode proposée aux séries chronologiques qualitatives en tableaux croisés de dénombrements est illustrée par le problème de la prédiction de la mortalité due au cancer.

MOTS CLÉS: Modèles à espace d'états, modèles linéaires généralisés, filtre de Kalman.

1. INTRODUCTION

Le problème de la modélisation et de la projection des séries chronologiques qualitatives en tableaux croisés est assez fréquent dans le cas de la planification et de la prise de décisions. Les données se présentent généralement sous la forme d'une série assez longue de tableaux multidimensionnels de comptes basés sur un grand nombre d'observations recueillies à des intervalles réguliers. Ainsi, la série des données sur la mortalité due au cancer au Canada représente les données annuelles pour chaque province, regroupées selon le siège du cancer, l'âge et le sexe (voir exemple à la partie 5). On obtient les séries de la mortalité de sources administratives avec un retard d'environ deux années avant la publication des données. Le problème des délais préoccupe beaucoup les utilisateurs et les chercheurs, et il est évident qu'il serait très utile de projeter ces séries de données au moins jusqu'à l'année courante avant leur publication. Pour cela, on pourrait considérer que la nature inhérente des données serait stochastique (Brillinger, 1986) en dépit de leur origine de sources administratives. On peut alors raisonnablement supposer qu'il existe une dépendance sérielle des séries en raison de certains facteurs communs, connus et inconnus. Si les données étaient normales, on pourrait utiliser diverses méthodes de séries chronologiques bien connues, comme par exemple celles exposées dans les ouvrages classiques de Box et Jenkins (1970), de Fuller (1976) et de Harvey (1981). Toutefois, dans le cas des données non normales, comme par exemple celles suivant une loi de Poisson tels les comptes de la mortalité due au cancer, il faut considérer d'autres méthodes de séries chronologiques.

Il existe une littérature considérable consacrée à l'analyse des données non normales recueillies dans le temps. En particulier, pour des événements qualitatifs répétés, Koch, Landis, Freeman, Freeman et Lehnen (1977) utilisent les moindres carrés généralisés pour ajuster des modèles non linéaires dans lesquels on considère que le temps est un autre facteur de classification. Les travaux de Stiratelli, Laird et Ware (1984) décrivent une famille de modèles mixtes qui conviennent à des réponses dichotomiques répétées, dans lesquelles on pose certaines hypothèses sur les structures de covariance. Zeger, Liang et Self (1985), par contre, considèrent les modèles de régression logistique pour des observations binaires répétées dans une dépendance temporelle autorégressive de premier ordre simple. Stram, Wei et Ware (1988) considèrent des méthodes de modélisation des événements qualitatifs ordonnés dans le temps, méthodes dans lesquelles on suppose que les paramètres des modèles sont spécifiques à chaque moment ou point dans le temps et sont estimés par la maximisation des vraisemblances spécifiques au moment. La normalité asymptotique conjointe de ces estimations sert à caractériser la dépendance entre des observations répétées. Les travaux de Morton (1987) et de Preisler (1989) portent sur l'ajustement de modèles linéaires généralisés avec effets aléatoires emboîtés dans les effets jour/temps aléatoires. Ces communications, par contre, ne portent pas sur le problème de la projection que l'on examine dans le présent article.

Selon Cox (1981), les différentes approches des séries chronologiques aux données non normales peuvent être classées en deux catégories, à savoir les modèles à observations et les modèles à paramètres. Certaines méthodes se rattachant au premier type sont dues à Kalbfleisch et Lawless (1984, 1985) et à Kaufmann (1987), où l'on considère des modèles de Markov pour la régression (ou des probabilités de transition) avec des résultats qualitatifs. On peut également voir Zeger et Qaqish (1988) pour une approche de quasi-vraisemblance aux modèles de régression de Markov pour des séries chronologiques générales. Smith et Brunsdon (1989) ont récemment proposé une autre méthode dans laquelle on suppose la normalité approximative après transformation additive- logistique multivariée des données multinomiales d'abord, puis l'utilisation de modèles ARMM (auto-régressifs de moyenne mobile). Certaines méthodes se rattachant au deuxième type, c'est-à-dire les modèles à paramètres, sont dues à West, Harrison et Migon (1985), celles-là avec une structure bayesienne pour l'extension dynamique de modèles linéaires généralisés, à Kitagawa (1987) qui offre une meilleure approche

[1] A.C. Singh et G.R. Roberts, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6

Bates, D.V. & Sizto, R. (1987). Air pollution and hospital admissions in Southern Ontario: the acid summer haze effects. *Environmental Research* **43**, 317-331.

Breslow, N.E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics* 33, 38-44.

Brillinger, D.R. & Preisler, H.K. (1983). Maximum likelihood estimation in a latent variable problem. In *Studies in Econometrics, Time Series and Multivariate Statistics* (S. Karlin, T. Amemiya, L.A. Goodman eds.) Academic Press, New York, pp. 31-65.

Cox, D.R. (1981). Statistical analysis of time series, some recent developments. *Scandanivan Journal of Statistics* 8, 93-115.

Dean, C. & Lawless, J.R. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84**, 467-471.

Dean, C., Lawless, J.R. & Willmot, G.E. (1989). A mixed Poisson-inverse-Gaussian regresion model. *Canadian Journal of Statistics* **17**, 171-181.

Hinde J. (1982). Compounded Poisson regression models. In: GLIM 82 (R. Gilchrist, ed.). Springer, New York, pp. 109-121.

Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* 15, 209-225.

Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

McCullagh, P. & Nelder, J.P. (1983). *Generalized Linear Models*. Chapman and Hall, New York.

Morton, R. (1987). A generalized linear model with strata of variation. *Biometrika* **74**, 247-257.

Rao, C.R. (1973). *Linear Statistical Inference and its Application*. Wiley, New York.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

devenir complexe pour les très longues séries. Pour contourner cette difficulté, nous utilisons une matrice de covariances provisoire qui peut être inversée algébriquement lorsqu'on estime β au moyen de l'équation (3.2). Nous calculons ensuite la matrice de dispersion du vecteur des paramètres de régression estimés en utilisant la covariance réelle. Zeger (1988) s'est servi de cette méthode pour une série de données d'énumération. Considérons tout d'abord une approximation $\tilde{V}_i$ de V_i, qui est définie par l'équation

$$\tilde{V}_i = \tilde{\Lambda}_i \Omega_i \tilde{\Lambda}_i + \tau \lambda_i \lambda_i^T, \tag{4.1}$$

où $\tilde{\Lambda}_i = \mathrm{diag}\{(\lambda_{i1} + \phi(\Gamma + 1)\lambda_{i1}^2)^{1/2}, \ldots, (\lambda_{in_i} + \phi(\Gamma + 1)\lambda_{in_i}^2)^{1/2}\}$. Notons que $\tilde{V}_i$ a les mêmes éléments diagonaux que V_i. Si nous posons $G_i = \tilde{\Lambda}_i \Omega_i \tilde{\Lambda}_i$, le théorème réciproque du binôme pour les matrices (Rao, 1973, p.33) implique que

$$\tilde{V}_i^{-1} = G_i^{-1}\left[I - \tau \lambda_i \lambda_i^T G_i^{-1}\left(1 + \tau \lambda_i^T G_i^{-1} \lambda_i\right)^{-1}\right], \tag{4.2}$$

où $G_i^{-1} = \tilde{\Lambda}_i^{-1} \Omega_i^{-1} \tilde{\Lambda}_i^{-1}$. Dans le cas d'une structure d'autocorrélation autorégressive, nous pouvons calculer explicitement Ω_i^{-1} (Zeger, 1988). Si nous utilisons la matrice des covariances provisoire pour estimer la covariance des paramètres de régression estimés, nous avons la formule suivante

$$\mathrm{Cov}\,(\tilde{\beta}) = \tilde{H}^{-1}\left(\sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} V_i \tilde{V}_i^{-1} D_i\right)\tilde{H}^{-1} \tag{4.3}$$

où $\tilde{H} = \sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} D_i$. De cette façon, V_i n'a pas besoin d'être inversée.

Pour de très longues séries chronologiques, comme celles qui renferment plus de 100 observations, nous proposons une formule encore plus simple pour la matrice des covariances provisoire. En posant $\alpha = 0$, nous obtenons une matrice des covariances provisoire désignée par Λ_i que l'on peut inverser facilement. De plus, comme la valeur estimée de β ne dépend pas de α dans les circonstances, il n'est pas nécessaire d'exécuter une itération conjointe pour les valeurs estimées de β et de α.

5. ANALYSE

Nous venons de décrire des modèles de régression pour des séries chronologiques parallèles de données d'énumération. Ce genre de modèles est utilisé dans les études que réalise couramment la Direction générale de la protection de la santé sur les effets de la pollution atmosphérique sur la santé. Dans ces études, on essaie d'établir un rapport entre le nombre de personnes admises quotidiennement à l'hôpital pour des troubles respiratoires et niveau de pollution atmosphérique enregistré à chaque jour dans le voisinage de chaque hôpital. Trois sources de variation sont considérées: la variation du nombre d'admissions entre hôpitaux; la variation des observations pour chaque hôpital; et la corrélation des observations dans le temps.

La dispersion des observations est décrite par un modèle mixte à effets aléatoires où l'on suppose que l'espérance conditionnelle est égale à la variance conditionnelle, étant donné les effets aléatoires. Puisque seuls les deux premiers moments des observations conditionnelles sont définis, on a recours à des équations d'estimation pour estimer les paramètres de régression et de dispersion. On réussit à obtenir des estimateurs convergents de ces paramètres ainsi qu'un estimateur convergent de la variance des paramètres de régression. Toutefois, il n'est pas possible d'obtenir un estimateur de la variance des paramètres de dispersion à cause d'un manque de renseignements sur les moments d'ordres supérieurs. Qu'à cela ne tienne, si l'analyse est centrée sur les paramètres de régression et qu'elle traite les paramètres de dispersion comme des facteurs dérangeants, la lacune évoquée ci-dessus n'a pas de conséquences sérieuses dans la pratique.

6. BIBLIOGRAPHIE

Anderson, D.A. & Hinde, J. (1988). Random effects in generalized linear models and the EM algorithm. *Communication in Statistics* 17, 3847-3856.

On détermine un estimateur de moment de ϕ et de ρ_ℓ en considérant que ϕ représente le degré de dispersion des observations d'une série donnée et ρ_ℓ, le coefficient d'autocorrélation avec décalage ℓ. Un estimateur convergent ($n_i \to \infty$) de la variance conditionnelle moyenne $\sum_{t=1}^{n_i} \mathrm{Var}(Y_{it} | \eta_i)/n_i$ dans la série i est défini

$$\hat{v} = \sum_{t=1}^{n_i} \left(y_{it} - \hat{\eta}_i \hat{\lambda}_{it} \right)^2 / n_i \tag{3.4}$$

où $\hat{\lambda}_{it} = \exp(X_{it}^T \hat{\beta})$, et

$$\hat{\eta}_i = \left(\sum_{t=1}^{n_i} y_{it} \right) \left(\sum_{t=1}^{n_i} \hat{\lambda}_{it} \right)^{-1} \tag{3.5}$$

est un estimateur convergent de η_i. L'équation (2.3) nous permet de déduire qu'un estimateur convergent $\hat{\phi}$ de ϕ, lorsque $N \to \infty$, est défini par l'expression

$$\hat{\phi} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{n_i} \left[\left(y_{it} - \hat{\eta}_i \hat{\lambda}_{it} \right)^2 - \hat{\eta}_i \hat{\lambda}_{it} \right]}{\sum_{i=1}^{N} \sum_{t=1}^{n_i} \left(\hat{\eta}_i \hat{\lambda}_{it} \right)^2}. \tag{3.6}$$

On obtient un estimateur de ρ_ℓ en posant un estimateur de moment du coefficient d'autocorrélation avec décalage ℓ égal à son espérance mathématique, étant donné η_i, laquelle espérance est déduite de l'équation (2.3); ainsi,

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{N} \sum_{t=\ell+1}^{n_i} \left(y_{it} - \hat{\eta}_i \hat{\lambda}_{it} \right) \left(y_{i,t-\ell} - \hat{\eta}_i \hat{\lambda}_{i,t-\ell} \right)}{\hat{\phi} \sum_{i=1}^{N} \sum_{t=\ell+1}^{n_i} \hat{\eta}_i^2 \hat{\lambda}_{it} \hat{\lambda}_{i,t-\ell}} \tag{3.7}$$

pour $\ell = 1,\ldots,k$. L'estimateur ci-dessus permet d'estimer les paramètres de corrélation d'un processus autorégressif à l'aide des équations de Yule-Walker (Zeger, 1988). Dans le cas d'un processus autorégressif du premier degré, c'est le coefficient d'autocorrélation avec décalage un ($\hat{\rho}_1$) qui fournit une estimation du paramètre de corrélation.

Enfin, on obtient un estimateur de moment de τ en posant l'estimateur de moment de la variance inconditionnelle égal à son espérance mathématique, ce qui donne

$$\hat{\tau} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{n_i} \left[\left(y_{it} - \hat{\lambda}_{it} \right)^2 - \hat{\lambda}_{it} \left(1 + \hat{\phi} \hat{\lambda}_{it} \right) \right]}{\left(\hat{\phi} + 1 \right) \sum_{i=1}^{N} \sum_{t=1}^{n_i} \hat{\lambda}_{it}^2}. \tag{3.8}$$

On termine le processus d'estimation en révisant les valeurs estimées de β, définies en (3.2), et de α, définies par les équations (3.6) à (3.8), jusqu'à ce qu'il y ait convergence. La matrice des covariances estimée de $\hat{\beta}$ est $\mathrm{Cov}(\hat{\beta}) = H(\hat{\beta},\hat{\alpha})^{-1}$. Notons que la covariance de $\hat{\beta}$ ne dépend aucunement de la variance de $\hat{\alpha}$ à cause du caractère indépendant de l'espérance inconditionnelle λ_{it} et des paramètres de dispersion α.

Puisque seuls les deux premiers moments η_i ou de ε_{it} ont été définis, il n'est pas possible habituellement d'établir l'erreur estimée pour $\hat{\alpha}$. Heureusement, notre analyse est centrée sur les paramètres de régressin β et elle traite les paramètres de dispersion comme des facteurs dérangeants. Une façon de choisir un modèle de dispersion convenable serait d'effectuer une analyse de sensibilité pour évaluer l'effet de la forme d'α sur $\mathrm{Cov}(\hat{\beta})$. Par exemple, une valeur $\tau = 0$ indique que les séries parallèles peuvent être envisagées comme une seule grande série chronologique. Si $\tau = \rho = 0$ et $\phi > 0$, alors les observations ne représentent qu'une série de données d'énumération non corrélées disparates. Dean et coll. (1989) examinent d'autres méthodes pour évaluer le degré de dispersion des observations d'une série chronologique.

4. ESTIMATION POUR DE LONGUES SÉRIES CHRONOLOGIQUES

Beaucoup d'applications produisent de longues séries chronologiques. Les méthodes que nous utilisons dans ces circonstances exigent une inversion répétée de la matrice des variances-covariances V_i, opération qui peut et qui est indépendant de ϵ_{it}. D'après Zeger (1988), nous supposons que l'espérance et la variance conditionnelles des observations sont définies

$$E(y_{it}|\eta_i,\ \epsilon_{it}) = \mathrm{Var}(y_{it}|\eta_i,\ \epsilon_{it}) = \eta_i\ \epsilon_{it}\ \lambda_{it}, \tag{2.2}$$

où $\lambda_{it} = \exp(x_{it}^T\beta)$. La covariance conditionnelle de deux observations quelconques de la même série est supposée nulle. La moyenne, la variance et la covariance de deux observations de la même série, étant donné η_i, s'écrivent donc

$$\begin{aligned} E(y_{it}|\eta_i) &= \eta_i\lambda_{it} \\ \mathrm{Var}(y_{it}|\eta_i) &= \eta_i\lambda_{it} + \phi\eta_i^2\lambda_{it}^2, \quad \text{et} \\ \mathrm{Cov}(y_{it},\ y_{i,t+\ell}|\eta_i) &= \phi\rho_\ell\eta_i^2\lambda_{it}\lambda_{i,t+\ell}. \end{aligned} \tag{2.3}$$

La moyenne, la variance et la covariance inconditionnelles sont définies

$$\begin{aligned} E(y_{it}) &= \lambda_{it} \\ \mathrm{Var}(y_{it}) &= \lambda_{it} + (\tau + \phi[\tau+1])\ \lambda_{it}^2, \quad \text{et} \\ \mathrm{Cov}(y_{it},\ y_{i,t+\ell}) &= (\tau + \phi[\tau+1]\rho_\ell)\,\lambda_{it}\ \lambda_{i,t+\ell}. \end{aligned} \tag{2.4}$$

La matrice des variances-covariances pour la série i est

$$\mathrm{Cov}(\mathrm{Y}_i) = \Lambda_i + \Lambda_i\mathrm{R}_i(\alpha)\Lambda_i \equiv \mathrm{V}_i, \tag{2.5}$$

où $\mathrm{Y}_i = (y_{i1},\ldots,y_{in_i})^T$, $\Lambda_i =$ *diag* $(\lambda_{i1},\ldots,\lambda_{in_i})$, $\alpha = (\phi,\tau,\rho_1,\ldots,\rho_k)$ et

$$\mathrm{R}_i(\alpha) = \tau\mathrm{J}_i + \phi(\tau+1)\Omega_i. \tag{2.6}$$

Dans l'équation ci-dessus, J_i est une matrice (n_i x n_i) formée de uns et Ω_i est une matrice de corrélation (n_i x n_i) dont les éléments de la ℓ-ième diagonale sont définis par ρ_ℓ. Notre objectif est d'estimer les paramètres de régresion β et les paramètres de dispersion α.

3. ESTIMATION DES PARAMÈTRES

Comme il n'existe aucune hypothèse sur la distribution des observations conditionnelles ou des variables aléatoires ξ_{it} et η_{it}, nous ne pouvons recourir à la méthode de vraisemblance pour estimer les paramètres. Toutefois, étant donné que les deux premiers moments des observations ont été définis, nous pouvons utiliser des équations d'estimation conçues pour les données périodiques (Liang et Zeger, 1986).

Étant donné un $N^{1/2}$-estimateur convergent $\hat{\alpha}$ de α, l'estimateur $\hat{\beta}$ du vecteur de régression β satisfait l'équation d'estimation

$$\mathrm{U}(\hat{\beta}|\hat{\alpha}) = \sum_{i=1}^{N} \mathrm{D}_i(\hat{\beta})^T\mathrm{V}_i^{-1}(\hat{\beta},\ \hat{\alpha})\left(\mathrm{Y}_i - \lambda_i(\hat{\beta})\right) = 0, \tag{3.1}$$

où $D_i(\beta) = \partial\lambda_i/\partial\beta = \Lambda_i X_i$, $\lambda_i = (\lambda_{i1},\ldots,\lambda_{in_i})^T$ et $X_i = (x_{i1},\ldots,x_{in_i})^T$. La valeur estimée $\hat{\beta}$ est établie au moyen d'une méthode itérative (Liang et Zeger, 1986). Étant donné la valeur estimée courante $\hat{\beta}^{(h)}$ de β et $\hat{\alpha}^{(h)}$ de α, on obtient la nouvelle valeur estimée $\hat{\beta}^{(h+1)}$ par

$$\hat{\beta}^{(h+1)} = \hat{\beta}^{(h)} + \mathrm{H}\left(\hat{\beta}^{(h)}, \hat{\alpha}^{(h)}\right)^{-1}\mathrm{U}\left(\hat{\beta}^{(h)}|\hat{\alpha}^{(h)}\right) \tag{3.2}$$

où

$$\mathrm{H} = -E(\partial\mathrm{U}/\partial\beta) = \sum_{i=1}^{N}\mathrm{D}_i^T\mathrm{V}_i^{-1}\mathrm{D}_i. \tag{3.3}$$

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MODÈLES DE RÉGRESSION POUR SÉRIES CHRONOLOGIQUES PARALLÈLES DE DONNÉES D'ÉNUMÉRATION

R. Burnett, D. Krewski et J. Shedden[1]

RÉSUMÉ

Dans cet article, nous considérons des modèles de régression pour des séries chronologiques parallèles de données d'énumération. Nous analysons plus particulièrement les effets de processus mixtes à effets aléatoires qui servent à traduire la variation des observations entre les séries et à l'intérieur des séries de même que la corrélation des observations dans le temps. Nous estimons les paramètres de régression et de dispersion au moyen d'équations d'estimation.

1. INTRODUCTION

Les modèles de régression pour données d'énumération sujettes à une forte variance ont fait l'objet de recherches soutenues ces dernières années (McCullagh et Nelder, 1983). Cox (1981) a étudié des modèles où la dispersion est proportionnelle à la variance des observations tandis que Breslow (1984), Morton (1987), Lawless (1987) et Dean et Lawless (1989) ont étudié des structures de variance inspirées de la distribution binomiale négative, qui se présentent comme une distribution de Poisson composée. Hinde (1982) et Dean et coll. (1989) se sont respectivement penchés sur la distribution de Poisson normale et la distribution de Poisson gaussienne inverse, tandis que Brillinger et Preisler (1983) se sont intéressés à des distributions de Poisson composées arbitrairement. Morton (1987) a étudié des modèles à effets aléatoires emboîtés pour données d'énumération en utilisant des méthodes de quasi-vraisemblance et Anderson et Hinde (1988) ont intégré ces modèles à la famille exponentielle en se servant de méthodes de vraisemblance et de l'algorithme EM. Zeger et coll. (1988) ont traité la même question sauf que dans ce cas, les effets aléatoires se rattachent à des covariables mesurées. Zeger (1988) et Zeger et Qaqish (1988) ont introduit l'autocorrélation dans les modèles pour une série simple de données d'énumération.

Dans cet article, nous nous intéressons particulièrement aux modèles de régression pour des séries chronologiques parallèles de données d'énumération. On recueille de telles données lorsqu'on étudie, par exemple, l'incidence de la pollution de l'air ambiant sur le nombre de personnes admises quotidiennement à l'hôpital à cause de troubles respiratoires (Bates et Sizto, 1987). Comme l'étude porte normalement sur plusieurs hôpitaux, on dispose de plusieurs séries chronologiques pour l'analyse. De plus, les dossiers médicaux étant mis à jour au fil des ans, nous allons considérer des méthodes d'estimation qui peuvent être utilisées avec de longues séries chronologiques. Si l'estimation de paramètres se fait par des fonctions de vraisemblance, il faut normalement recourir à l'intégration numérique ou poser des hypothèses concernant le degré de dispersion (Zeger et coll., 1988). Pour de longues séries chronologiques, l'intégration numérique peut s'avérer une opération complexe. Or, Zeger (1988) considère une série chronologique simple de données d'énumération et calcule la valeur estimée des paramètres de régression et de dispersion correspondants à l'aide d'équations d'estimation. Cette méthode n'exige pas de fixer une limite supérieure pour le degré de dispersion ni de recourir à l'intégration numérique. Dans cet article, nous appliquons aussi cette méthode aux séries chronologiques multiples.

2. DÉFINITION DU MODÈLE

Soit y_{it} l'observation tirée de la série i au passage t $(t = 1,\dots,n_i;\ i = 1,\dots,N)$. Bien que nous supposions que les observations s'échelonnent à intervalles réguliers, notre analyse n'exclut pas la possibilité de données manquantes. Soit x_{it} un vecteur $(p \times 1)$ de covariables et le vecteur des paramètres de régression correspondant $\beta = (\beta_1,\dots,\beta_p)^T$. De plus, soit ϵ_{it} une variable aléatoire strictement positive ayant une espérance mathématique égale à un et une covariance définie par l'équation

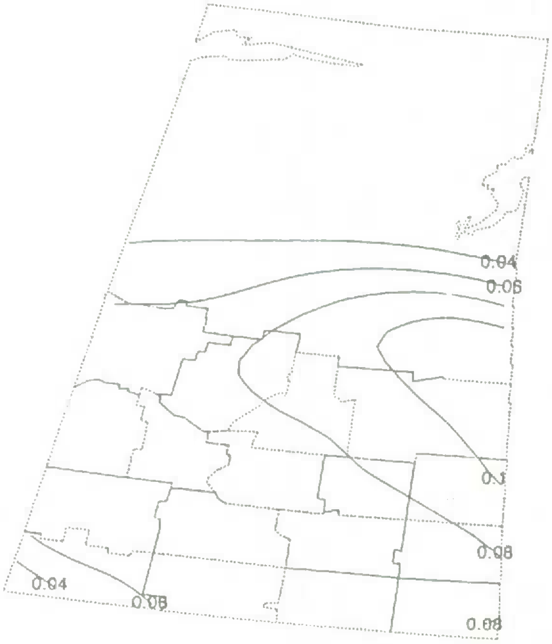
$$\mathrm{Cov}(\epsilon_{it},\ \epsilon_{i,t+\ell}) = \phi\rho_\ell, \tag{2.1}$$

où $\Phi>0$ et $|\rho_\ell|<1$ représente le coefficient d'autocorrélation avec décalage $\ell = 1,2,\dots,\ k\leq \max(n_i)$.

Désignons par η_i l'effet aléatoire pour la série i, ayant une espérance mathématique de un et une variance $\tau>0$,

[1] Centre d'hygiène du milieu, Santé et Bien-être social Canada, Ottawa, (Ontario), Canada.

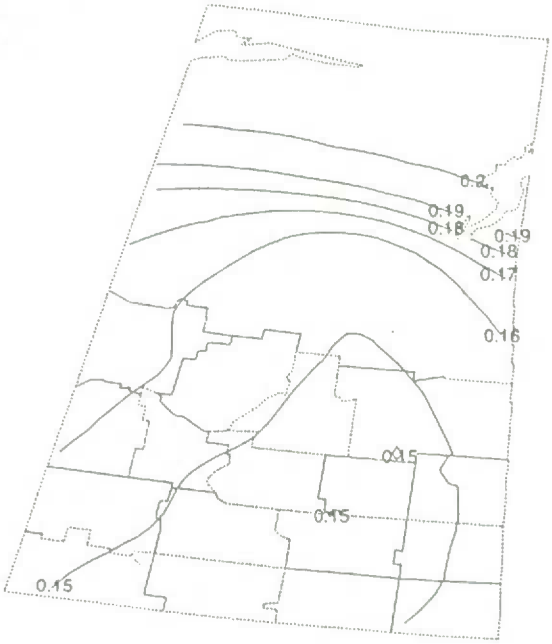
EFFET DES JOURS OUVRABLES



Modèle de Poisson avec effet des jours ouvrables

Figure 7

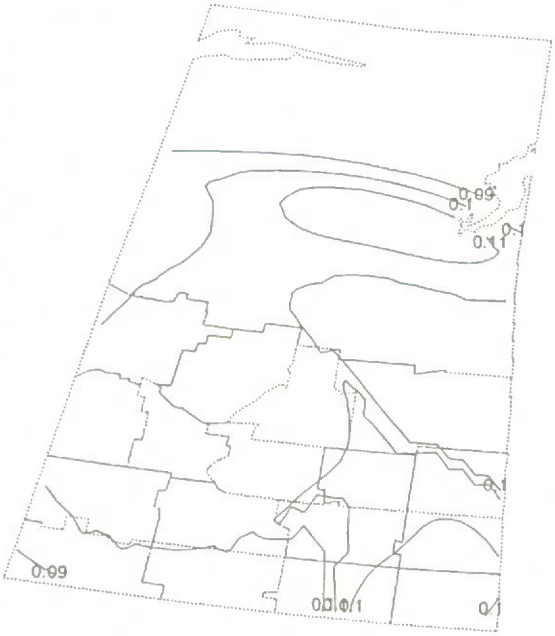
TAUX ANNUELS DE NATALITÉ



Modèle normal logarithmique de Poisson

Figure 8

VALEURS ESTIMÉES SIGMA



Modèle normal logarithmique de Poisson

Figure 9

TAUX ANNUELS DE NATALITÉ

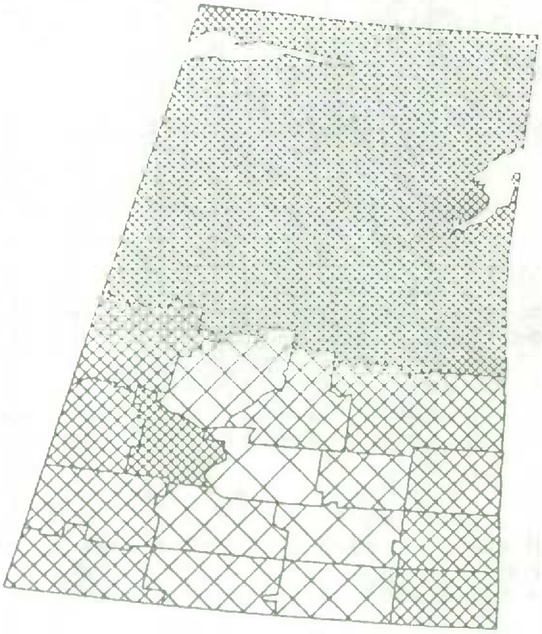


Figure 3

POIDS DES DIVISIONS DE RECENSEMENT

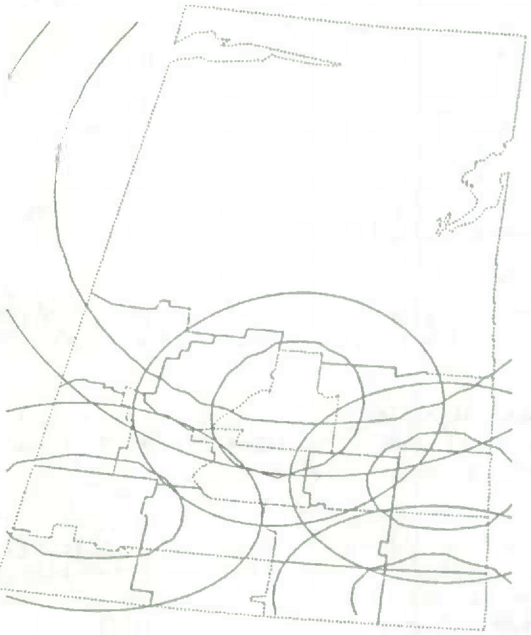
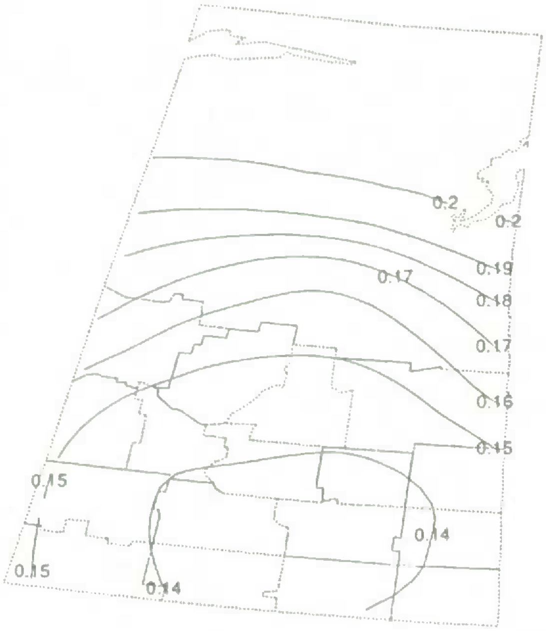


Figure 4

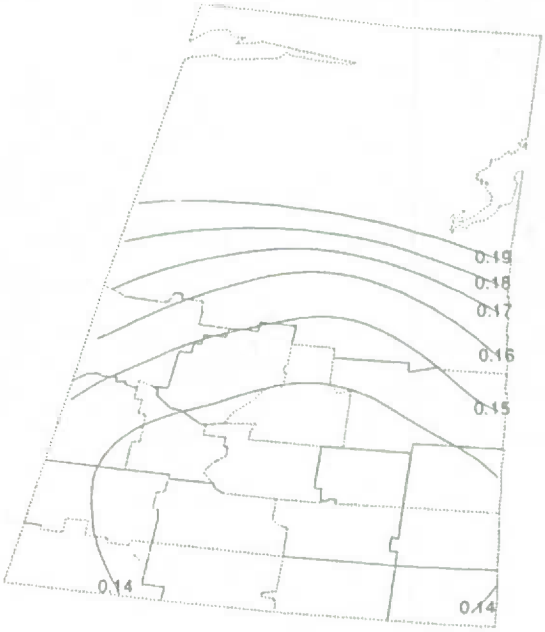
TAUX ANNUELS DE NATALITÉ



Modèle de Poisson ordinaire

Figure 5

TAUX ANNUELS DE NATALITÉ



Modèle de Poisson avec effet des jours ouvrables

Figure 6

LÉGENDES DES FIGURES

Figure 1. Nombre total de naissances chez les femmes de 25 à 29 ans en 1986 pour les 18 divisions de recensement de la Saskatchewan, ainsi que le nombre de femmes recensées dans ce groupe d'âge le 3 juin de la même année. (Tel que mentionné dans le texte, le dernier chiffre de chaque total a été arrondi au nombre le plus près de 2 ou 7).

Figure 2. Taux annuels de natalité pour les 18 divisions de recensement pour les femmes de 25 à 29 ans.

Figure 3. Les taux de la figure 2 représentés au moyen d'une carte hachurée, la concentration des hachures étant proportionnelle au taux.

Figure 4. Les poids, $W_i(x,y)$ utilisés dans les équations (1) ou (2) calculés à l'aide de l'expression (4) pour quatre des divisions de recensement. Pour plus de clarté, les poids ne sont pas montrés pour toutes les divisions.

Figure 5. Représentation graphique de l'expression (5) utilisant les poids de l'expression (4), B_i étant le nombre de naissances enregistrées dans la division de recensement i et N_i le nombre correspondant de femmes recensées âgées de 25 à 29 ans.

Figure 6. Valeur estimée du taux de natalité en supposant que le nombre de naissances, B, étant donnée la population à risque, N, suit une distribution de Poisson avec moyenne N exp $\{\alpha \pm \beta\}$, le signe + s'appliquant aux jours ouvrables et le signe — aux jours non ouvrables. La méthode d'ajustement à pondération locale est utilisée pour obtenir la valeur estimée $\exp\{\hat{\alpha}(x,y)\}$.

Figure 7. Représentation de l'effet des jours ouvrables estimé $\hat{\beta}(x,y)$ obtenu selon la figure 6.

Figure 8. Représentation comparable à la figure 6, sauf qu'une variable d'erreur normale est ajoutée au modèle de prédiction linéaire.

Figure 9. Représentation comparable à la figure 7, sauf (comme dans la figure 8) qu'une variable d'erreur normale a été ajoutée au modèle de prédiction linéaire.

NOMBRE DES NAISSANCES ET POPULATION

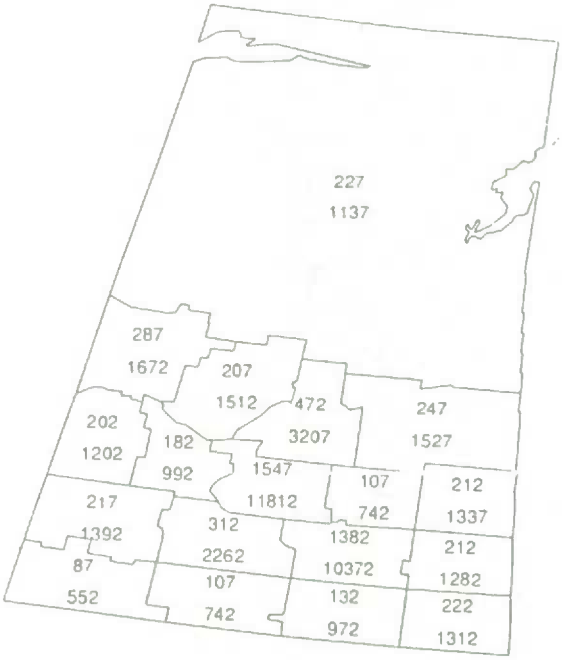


Figure 1

TAUX ANNUELS DE NATALITÉ

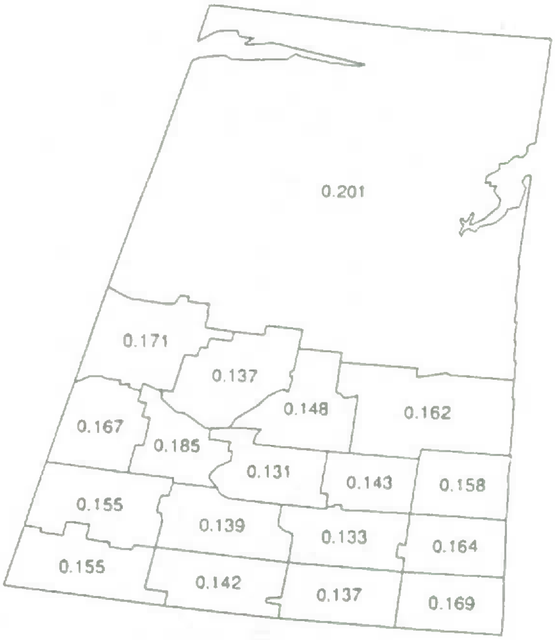


Figure 2

BIBLIOGRAPHIE

Bock, R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika 35, 179-197.

Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics. Biometrics 42, 693-734.

Brillinger, D. Spatial-temporal modelling of spatially aggregate birth data. R. (1990).

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83, 596-610.

Cleveland, W.S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. Technometrics 17, 447-454.

Cliff, A.D. and Ord, J.K. (1975). Model building and the analysis of spatial pattern in human geography. J. Royal Stat. Soc. 37, 297-348.

Dyn, N. and Wahba, G. (1982). On the estimation of functions of several variables from aggregated data. SIAM J. Math. Anal. 13, 134-152.

Franke, R. (1982). Scattered data interpolation: tests of some methods. Math. Comp. 38, 181-200.

Gilchrist, W.G. (1967). Methods of estimation involving discounting. J. Royal. Stat. Soc. 29, 355-369.

Manton, K.G., Woodbury, M.A., Stallard, E., Riggan, W.B., Creason, J.P. and Pelom, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. J. Amer. Statist. Assoc. 84, 637-650.

Miyaoka, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. Canadian J. Stat. 17, 123-140.

Pierce, D.A. and Sands, B.R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Dept., Oregon State University.

Preparata, F.P. and Shamos, I. (1985). Computational Geometry. Springer, New York.

Shaban, S.A. (1988). Poisson log-normal distributions. Pp. 195-210 in Lognormal Distributions (eds. E.L. Crow and K. Shimizu). M. Dekker, New York.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. J. Amer. Statist. Assoc. 84, 276-283.

Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. J. Amer. Statist. Assoc. 74, 519-536.

Tsutakawa, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. J. Amer. Statist. Assoc. 83, 37-42.

Tukey, J.W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Included in The Collected Works of J.W. Tukey, Vol. 5 (1988), (Ed. W.S. Cleveland). Wadsworth, Pacific Grove.

ANNEXE

Cette annexe contient quelques détails sur les calculs. Les limites de la province et des divisions de recensement forment des polygones. Pour calculer les poids $w_i(x,y)$, il a fallu utiliser un programme qui vérifiait si un point donné se trouvait dans un polygone donné. Pour calculer la moyenne et la variance d'un point aléatoire à l'intérieur d'un polygone donné, il a fallu recourir à une opération par laquelle le polygone était divisé en triangles. Ces programmes sont analysés dans Preparata et Shamos (1985). La fonction de vraisemblance a été maximisée au moyen du programme FORTRAN va09a de Harwell. Pour le calcul en parallèle, la grille de 40 par 40 a été décomposée en 20 segments disjoints.

variables dans les circonstances. Nous allons donc supposer que les variables absentes du modèle sont représentées collectivement par une variable d'erreur. Nous allons aussi supposer que, étant donné ε_i, la variable aléatoire B_i suit une distribution de Poisson de moyenne $N_i\mu \exp\{\varepsilon_i\}$ et que ε_i suit une distribution normale de moyenne 0 et de variance σ^2. La variable aléatoire B suit donc une distribution normale logarithmique de Poisson. Pour plus de renseignements sur cette distribution, veuillez vous référer à Shaban (1988).

Un inconvénient majeur du modèle normal logarithmique de Poisson est qu'il n'existe pas d'expression analytique pour la loi de probabilité. Cependant, le modèle se prête très bien à l'introduction d'effets et au traitement de variables manquantes. Les travaux de Bock et Lieberman (1970) et de Pierce et Sands (1975) nous indiquent que l'on peut résoudre la difficulté évoquée ci-dessus en recourant à l'intégration numérique. Nous pouvons exprimer la fonction de probabilité par la formule

$$p(y) = \frac{1}{y!} \int (ve^{\sigma z})^y \exp\{-ve^{\sigma z}\} \phi(z)dz$$

où ϕ est la densité normale standard, y correspond à B et v correspond à $N\mu$. Pour l'approximation de l'intégrale, on utilise un nombre fini de termes comprenant des noeuds et des poids.

Les figures 8 et 9 illustrent les résultats de l'ajustement fait à l'aide de 61 noeuds. La figure 8 montre des cercles vaguement concentriques autour des régions urbaines, comme dans les figures 5 et 6. La forme irrégulière des courbes de la figure 8 s'explique peut-être par la possibilité que le processus d'estimation ait convergé vers un extrémum local. La figure 9 n'est pas facile à décrire. Elle donne à croire que l'estimation est passablement variable. La valeur estimée $\hat{\sigma}$ se situe autour de 0.1 et est donc comparable à l'effet des jours ouvrables défini dans la section 6.

8. ANALYSE

L'analyse à pondération locale et les modèles d'effets aléatoires semblent être deux moyens de résoudre avec beaucoup de souplesse toute une série de problèmes ayant trait à des données géographiques. Les termes d'effet aléatoire ont deux fonctions importantes: traiter les effets manquants et renforcer le modèle de manière à produire de meilleures estimations pour les principaux paramètres. En ce qui concerne le modèle de Poisson ordinaire, les totaux élémentaires sont efficients mais il existe une variation extra-Poisson dans le cas qui nous occupe à cause des variables manquantes. La méthode exige beaucoup de temps d'ordinateur parce que l'intégration numérique et l'estimation par le maximum de vraisemblance sont exécutées à de nombreuses positions sur une grille; toutefois, les opérations se sont bien déroulées sur le système Sun 3/50 de Berkeley.

Beaucoup de recherches restent à faire; mentionnons au passage des sujets comme l'évaluation de l'ajustement, le calcul de la variabilité, le choix de la fonction de poids (y compris le choix de τ dans l'équation (4)), les analyses pour d'autres groupes d'âge et d'autres provinces et la définition d'asymptotes appropriées. D'autres résultats de recherches sont rapportés dans Brillinger (1990).

Parmi les articles récents portant sur l'analyse des données démographiques, notons ceux de Clayton et Kaldor (1987), de Tsutakawa (1988) et de Manton et coll. (1989). Contrairement à la présente étude, ces articles n'ont pas pour objet d'étudier la question des surfaces lisses.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à G. Brackstone, R. Gussela, R. Raby, B. Sander, P. Spector, M. Subhani et R. Villani pour lui avoir procuré les données et les cartes et l'avoir assisté dans la géométrie computationnelle et le calcul en parallèle.

5. MODÈLE DE POISSON ORDINAIRE

Dans cette analyse, nous nous intéressons uniquement aux femmes de 25 à 29 ans et aux naissances enregistrées chez ce groupe de femmes. Désignons par $i = 1, ..., 18$ la division de recensement et par N_i le nombre de femmes recensées dans la division i (il s'agit des chiffres du recensement du 3 juin 1986). Soit B_i le nombre total de naissances enregistrées chez les femmes de 25 à 29 ans en 1986.

Supposons que la loi de probabilité $p(.)$ de B_i est une distribution de Poisson de moyenne $N_i\mu$, étant le taux de natalité. Cette hypothèse repose sur l'idée que les anniversaires de naissance sont aléatoires (voir Brillinger, 1986).

Étant donné l'hypothèse du modèle de Poisson, l'estimation (à pondération locale) du taux de natalité à la position (x,y) est

$$\hat{\mu}(x,y) = \sum_i w_i(x,y)\ B_i \ / \sum_i w_i(x,y)\ N_i \qquad (5)$$

Ces valeurs sont calculées pour (x,y) sur une grille de 40 par 40. Le tracé de contours correspondant est reproduit dans la figure 5. On observe une progression lente de la valeur des contours. Le taux varie de .14 à .20, les valeurs les plus élevées étant observées dans la partie septentrionale de la province et les moins élevées étant concentrées dans les régions les plus urbanisées.

6. MODÈLE DE POISSON AVEC EFFET DES JOURS OUVRABLES

Bien que notre analyse soit essentiellement de nature spatiale, il est bon de s'arrêter brièvement à la dimension temporelle. Il est notoire que les taux de natalité varient selon les jours de la semaine à cause de la disponibilité des médecins (voir par exemple Miyaokoa, 1989). On ne peut donc s'attendre que le nombre total de naissances suive une distribution de Poisson homogène. Voici donc un modèle qui mérite d'être considéré. Soit j une variable indicatrice qui prend la valeur 1 si c'est pour un jour ouvrable et la valeur 2 si c'est pour un jour non ouvrable. Désignons par B_{ij} le nombre de naissances dans la division de recensement i.
Supposons que Bij suit une distribution de Poisson de moyenne $N_i \exp\{\alpha+\beta_j\}$. β_j représente l'effet des jours ouvrables et nous supposons que $\beta_1 + \beta_2 = 0$ pour que le modèle soit identifiable. S'il n'y a pas d'effet des jours ouvrables, alors $\beta_1,\ \beta_2 = 0$. Par ailleurs, en utilisant la méthode d'estimation à pondération locale décrite dans les sections 3 et 4, on peut estimer α et β comme des fonctions de la position.

La figure 6 donne la valeur estimée $\exp\{\hat{\alpha}(x,y)\}$ du taux annuel de natalité. Il est intéressant de constater que, par rapport à la figure précédente (modèle de Poisson ordinaire), les contours sont plus éloignés des régions urbaines. La figure 7 illustre l'effet des jours ouvrables estimé $\hat{\beta}_1(x,y)$. Dans ce cas, on observe une poussée dans l'est de la province. Les valeurs de $\hat{\beta}$ vont de .00 à .10 tandis que celles de $\hat{\alpha}$ vont de -2.0 à -1.6.

L'analyse que nous venons de faire donne à penser que des variables fondamentales peuvent influer sur les taux de natalité et que nous devons en tenir compte dans la modélisation et l'analyse.

7. MODÈLE NORMAL LOGARITHMIQUE DE POISSON

Étant donné une variable explicative multidimensionnelle x_i, un modèle de Poisson de moyenne $N_i \exp\{(x_i\ \theta\}$ pour B_i pourrait expliquer assez bien les données. Comme variables explicatives, pensons au régime alimentaire, au mode de vie, aux conditions atmosphériques, à l'environnement, aux jours fériés, au changement démographique, à la structure par âge, aux caprices des délimitations. Nous ne connaissons rien de ces moyennes mobiles, et celui de Stone (1977), qui porte plus particulièrement sur la régression. Dans son analyse de l'article de Stone, Brillinger (1977) propose l'équation (2) pour une distribution générale et justifie sa proposition en associant cette équation à une règle de Bayes. Cleveland et Devlin (1988) font un traitement très détaillé de la méthode des moindres carrés. Pour sa part, Staniswalis (1989) traite le cas p en général. Les avantages de la technique à pondération locale sont les suivants: aucune hypothèse "cachée" sur la distribution du modèle, mise en évidence des cas de non-additivité, variantes pour la résistance et l'influence, additivité simple des observations et aucune inversion de matrice (comme l'exige le krigeage par exemple).

4. CONSTRUCTION DES POIDS

Les données qui nous intéressent ici consistent essentiellement en des totaux pour des divisions de recensement. Nous ne pouvons donc pas utiliser directement la méthode décrite dans la section précédente. Il s'agit ici de déterminer des poids $w_i(x,y)$ qui traduisent convenablement l'effet de la division de recensement i sur la position (x,y). Supposons que $|R_i|$ désigne la superficie de la division de recensement i. Alors, la fonction de poids élémentaire est

$$w_i(x,y) = 1/|R_i| \quad \text{for} \quad (x,y) \quad \text{in} \quad R_i$$

et 0 dans le cas contraire. Dans cette étude, nous allons utiliser des fonctions fondamentales comme

$$w_i(x,y) = \frac{1}{|R_i|} \int_{R_i} W(x-u,y-v)dudv \tag{3}$$

où $W(.)$ est un noyau qui convient à des données non agrégées comme celles analysées dans Cleveland et Devlin (1988). Pour justifier l'utilisation de l'équation (3), nous pouvons considérer un processus ponctuel de Poisson. Les estimations seront calculées à l'aide des formules (1) ou (2), W_i étant remplacé par w_i.

Les poids particuliers utilisés à $r = (x,y)$ sont

$$w_i(r) = \exp\{-(1-p)^2 \| r - r_i \|^2 / 2\tau^2\} \tag{4}$$

à l'extérieur de l'ellipse $(r_0 - \bar{r}_i)S^{-1}{}_i(r_0 - r_i)' = d_0^2 = 5.991$ et 1 à l'intérieur. Dans l'équation (4),

$\|r\|^2 = x^2 = y^2$, $p = d_0 / \sqrt{(r - \bar{r}_i)S^{-1}{}_i(r - \bar{r}_i)'}$ et $\tau = .025$, tandis que $r_i = E\ U_i$ et $S_i = \text{var}\ U_i$,

où U_i est une variable aléatoire distribuée uniformément dans R_i. En clair, cela signifie que les divisions de recensement sont représentées par des ellipses ayant la même moyenne et la même matrice de variances-covariances. (Les valeurs exactes ont été déterminées après quelques essais pour faire en sorte que la surface de la première ellipse équivaille à environ 95% de la superficie de la division de recensement.)

La figure 4 illustre les contours à .50 et à .99 des poids $w_i(x,y)$ pour plusieurs divisions de recensement. On remarque que les contours suivent la forme générale des divisions.

Tobler (1979) et Dyn et Wahba (1982) décrivent d'autres fonctions de poids construites dans des circonstances semblables. La méthode exposée dans cette section présente quelques-uns des avantages de la technique à pondération locale: additivité des termes et absence d'interaction, aucune inversion de matrice requise et excellente résistance aux valeurs aberrantes.

Cliff et Ord (1975) (section 5.1) cherchent à mesurer l'influence que des comtés exercent l'un sur l'autre. Notre étude a plutôt pour but d'analyser l'influence d'un "comté" sur un point en particulier.

2. CARTES HACHURÉES

Méfiez-vous des cartes qui assignent une valeur moyenne à tout un territoire... (TRADUCTION)

Par ces mots, Tukey (1979) déplore l'utilisation de cartes comme celles représentées par les figures 2 et 3, où une valeur unique est associée à chaque division géographique. En fait, en examinant la figure 2, il est plus logique de penser que le taux de natalité ne varie pas aussi brusquement d'une division de recensement à l'autre. Un des objectifs de notre étude est justement d'établir des cartes qui décrivent une variation progressive des taux de natalité. Nous espérons que ces cartes permettront d'en arriver à des modèles généraux et favoriseront des analyses exploratoires indicatives.

Dans notre étude, nous nous intéressons aussi à la distribution statistique des fréquences proprement dites. Un modèle stochastique spécial qui s'impose naturellement dans les circonstances est le modèle de Poisson. Or, nous savons que les naissances sont liées à de nombreuses variables socio-économiques comme le régime alimentaire, le mode de vie, les conditions atmosphériques, l'environnement, les jours ouvrables, les jours fériés, la structure par âge. En outre, la population de la Saskatchewan a fluctué autour des chiffres du recensement tout le long de l'année 1986 et, finalement, l'âge des femmes visées par l'étude varie de 25 à 29 ans. En résumé, il semble que nous devions utiliser un modèle plus souple que celui de Poisson, un modèle qui pourrait tenir compte des covariables absentes. Nous utiliserons donc la distribution normale logarithmique de Poisson dans les circonstances. Comme le paramètre d'écart-type figure dans cette distribution, nous devons renforcer le modèle en combinant les valeurs d'observations.

3. ANALYSE À PONDÉRATION LOCALE

Dans le cas de données non agrégées, l'ajustement à pondération locale est une méthode appropriée pour estimer des quantités qui varient graduellement. Supposons que nous ayons une variable aléatoire Y avec une loi de probabilité $p(Y|\theta)$ qui dépend du paramètre de dimension fini θ. Supposons par ailleurs que nous voulions estimer θ pour la position définie par les coordonnées (x,y). Supposons que nous connaissions Y_i pour la position (x_i,y_i). Nous allons définir un poids $W_i(x,y)$ qui dépend de la distance entre (x_i,y_i) et (x,y).

Il s'agit d'estimer en maximisant la fonction de vraisemblance logarithmique pondérée

$$\sum_i W_i(x,y) \ \log \ p(Y_i|\theta) \tag{1}$$

ou (ce qui revient souvent au même) en résolvant le système d'équations d'estimation

$$\sum_i W_i(x,y) \ \psi(Y_i|\hat{\theta} = 0 \tag{2}$$

où $\psi(Y|\theta) = \partial \log p/\partial\theta$, la fonction score.

Afin d'illustrer la technique, prenons un cas simple, celui où Y suit une distribution normale de moyenne μ et de variance σ^2. On obtient la valeur estimée de μ (à pondération locale) en minimisant l'expression

$$\sum_i W_i(x,y) \ [Y_i - \mu]^2$$

ce qui donne

$$\hat{\mu}(x,y) = \sum_i W_i(x,y) \ Y_i/\sum_i W_i(x,y)$$

expression à première vue intéressante. Soulignons que ces formules sont souvent utilisées en infographie pour l'interpolation de données (voir, par exemple, Franke, 1982).

Parmi les textes traitant ce sujet, mentionnons l'article de Gilchrist (1967) sur l'"actualisation", celui de Pelto et coll. (1968) sur les moindres carrés, celui de Cleveland et Kleiner (1975), qui propose l'utilisation de

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

REPRÉSENTATION CARTOGRAPHIQUE DE DONNÉES AGRÉGÉES SUR LES NAISSANCES

D.R. Brillinger[1]

RÉSUMÉ

A l'aide de cartes de la province de Saskatchewan, nous faisons une analyse des naissances enregistrées en 1986 par division de recensement. Nous cherchons principalement à établir une relation entre le nombre de naissances et les régions géographiques; à cette fin, nous établissons des cartes en courbes de niveau qui décrivent le phénomène des naissances de façon uniforme. Une hiérarchie de modèles pour variables aléatoires quantitatives sont ajustés aux observations; il s'agit du modèle de Poisson, du modèle de Poisson avec effet des jours ouvrables et du modèle normal logarithmique de Poisson, l'utilisation de ce dernier étant justifiée par l'absence de covariables importantes dans l'analyse.

MOTS CLÉS: Données agrégées; établissement de cartes en courbes de niveau; variation extra-Poisson; analyse à pondération locale; cartes; distribution de Poisson; distribution normale logarithmique de Poisson; effets aléatoires; données géographiques; covariables non calculées.

1. INTRODUCTION

Dans cet article, nous nous intéressons à des données qui ont été agrégées en fonction de régions géographiques. Il devrait être facile d'analyser de telles données à cause des possibilités de représentation graphique (par exemple, on exprime le nombre en fonction de la région géographique, comme dans le cas de la représentation graphique de résidus, si souvent utilisée dans l'analyse de régression); toutefois dans le cas qui nous préoccupe, l'agrégation des données soulève des difficultés majeures.

Les données analysées ici concernent essentiellement le nombre quotidien de naissances chez les femmes de 25 à 29 ans pour l'année civile 1986 et pour chacune des 18 divisions de recensement de la Saskatchewan. Nous nous servons aussi des chiffres de population correspondants, établis lors du recensement de 1986, pour le calcul de taux. Nous avons choisi la Saskatchewan pour cette étude pilote parce que son territoire est modérément étendu et que les limites de ce territoire et des divisions de recensement sont régulières. (La seconde raison était importante au début de l'étude parce que nous ne disposions pas de cartes produites par ordinateur.) Nous avons choisi les femmes de 25 à 29 ans parce que c'est le groupe d'âge auquel correspond le plus grand nombre de naissances. Les données nous ont été fournies par Statistique Canada.

Les données analysées ont comme caractéristique d'être agrégées, non gaussiennes et non stationnaires dans l'espace et le temps.

Nous cherchons à établir un rapport entre le nombre de naissances et les régions géographiques et, plus particulièrement, à découvrir des régimes de fécondité selon les régions et, peut-être, des tendances inédites. L'étude comporte deux volets. Nous présentons tout d'abord une analyse à pondération locale de données agrégées; ensuite, nous définissons des modèles d'effets aléatoires, que nous ajustons en vue de traiter une variation extra-Poisson.

Nous tenons à préciser qu'il s'agit là d'un rapport préliminaire sur des recherches en cours. Par exemple, nous ne prenons pas en considération la structure très détaillée des données et ne disposons d'aucune mesure de la variance des diverses estimations. Nous nous concentrons surtout sur les totaux annuels pour les 18 divisions de recensement. Dans un autre ouvrage qui porte sur le même sujet (Brillinger (1990)), nous considérons aussi bien l'aspect temporel que géographique.

La Saskatchewan comprend 18 divisions de recensement. Ces divisions sont illustrées dans la figure 1. Les chiffres représentent le nombre total de naissances chez les femmes de 25 à 29 ans en 1986 ainsi que le nombre de femmes recensées le 3 juin de la même année. Nous voyons clairement que la partie septentrionale de la province est très peu peuplée. La figure 2 donne le taux annuel de natalité pour chaque division de recensement. Les divisions qui ont le taux de natalité le plus faible (.131 et .133 naissance par an) correspondent aux villes de Saskatoon et de Regina respectivement. La figure 3 représente les taux annuels de natalité au moyen d'une carte hachurée, la concentration des hachures étant proportionnelle au taux.

[1] D.R. Brillinger, Department of Statistics, University of California, Berkeley, California, U.S.A. 94720

SECTION 3

L'ANALYSE DES SÉRIES CHRONOLOGIQUES DE COMPTES

Cronkhite, F.R. (1986), "Use of Regression Techniques for Developping State and Area Employment and Unemployment Estimates," dans *Small Area Statistics: An International Symposium*, éds. R. Platek, J.N.K. Rao, C.E. Särndal et M.P. Singh, New York: John Wiley, pp. 160-174.

Drew, J.D., Singh, M.P. et Choudhry, G.H. (1982), "Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada," *Techniques d'enquête*, 8, 19-52.

Ericksen, E.P. (1974), "A Regression Method for Estimating Population of Local Areas," *Journal of the American Statistical Association*, 69, 867-875.

Fay, R.E. et Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Fuller, W.A. et Battese, G.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structures," Journal of the American Statistical Association, 68, 626-632.

Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates," dans *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 33-36.

Hartley, H.O. (1961), "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Function by Least Squares," *Technometrics*, 3, 269-280.

Henderson, C.R. (1975), "Best Linear Unbiased Estimation and Prediction under a Selection Model," *Biometrics*, 31, 423-447.

Jones, R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," *Journal of the Royal Statistical Society*, Sér. B, 42, 221-226.

Judge, G.G., Griffith, W.E., Hill, R.C., Lütkepohl, H. et Lee, T. (1985), *The Theory and Practice of Econometrics*, 2^{e} éd., New York, John Wiley.

Pantula, S.G. et Pollock, K.H. (1985), "Nested Analysis of Variance with Autocorrelated Errors," *Biometrics*, 41, 909-920.

Prasad, N.G.N. et Rao, J.N.K. (1990), "The Estimation of the Mean Square Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85 (sous presse).

Purcell, N.J. et Kish, L. (1980), "Post Censal Estimates for Local Areas (or Domains)," *International Statistical Review*, 48, 3-18.

Rao, J.N.K. (1986), "Synthetic Estimators, SPREE and Best Model-Based Predictors of Small Area Means," dans Série de monographies du Laboratoire de recherche en statistique et probabilités, N^{O} 97, Université Carleton, Ottawa.

Särndal, C.E. et Hidiroglou, M.A. (1989), "Small Domain Estimation: A Conditional Analysis," *Journal of the American Statistical Association*, 84, 266-275.

Scott, A.J., Smith, T.M.F. et Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," *International Statistical Review*, 45, 13-28.

Tiller, R. (1989), "A Kalman Filter Approach to Labor Force Estimation Using Survey Data," article présenté lors des réunions de l'American Statistical Association, Washington, D.C.

Tableau 1. Efficacité mensuelle moyenne du MPLNBE par rapport aux estimateurs synthétiques et à l'estimateur d'enquête, selon le modèle (7.1)

Petite région	$\bar{E}_{1i}$	$\bar{E}_{2i}$	$\bar{E}_{3i}$
1	3.56	1.14	10.11
2	2.86	1.92	8.10
3	2.89	1.63	8.19
4	2.67	3.59	7.56
5	2.87	3.94	8.13
6	3.01	2.87	8.56
7	3.07	0.82	8.72
8	2.98	0.94	8.52
9	3.44	2.05	9.74
10	3.08	1.64	9.72
11	3.24	1.85	9.18
12	2.98	6.45	8.43
13	2.75	1.88	7.80
14	2.98	2.16	8.50
15	3.14	1.98	8.89
16	2.73	4.91	7.74
17	2.76	2.72	7.83
18	2.81	3.37	8.02
19	2.95	4.11	8.34
20	3.43	1.16	9.78
21	3.14	1.44	8.94
Moyenne globale	3.02	2.50	8.61

8. CONCLUSIONS

On obtiendra le MPLNBE à l'aide du modèle généralisé de Fay-Herriot donné par (3.7), il faudra tout d'abord obtenir les estimateurs des paramètres σ^2, σ_v^2 et ρ d'après la méthode proposée par Pantula et Pollack (1985) puis remplacer chacune des variables correspondantes par son estimateur dans le MPLNB afin d'obtenir le MPLNBE. L'efficacité du MPLNBE sera évaluée selon la méthode exposée à la section 7 à l'aide des données de l'enquête sur la population active du Canada et d'une estimation de la matrice des covariances d'échantillonnage, $\sum$. Des travaux visant à obtenir une estimation de $\sum$ pour l'enquête sur la population active du Canada sont en cours.

On obtiendra aussi des approximations précises de l'erreur quadratique moyenne du MPLNBE et leurs estimateurs, d'après la méthode décrite dans Prasad et Rao (1990).

BIBLIOGRAPHIE

Battese, G.E., Fuller, W.A. et Harter, R. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Binder, D.A. et Dick, J.P. (1989), "Implications of Survey Designs for Estimating Seasonal ARIMA Models", Article présenté lors des réunions de l'American Statistical Association, Washington, D.C.

Brackstone, G.J. (1986), "Small Area Data: Policy Issues and Technical Challenges," dans *Small Area Statistics: An International Symposium*, éds. R. Platek, J.N.K. Rao, C.E. Särndal et M.P. Singh, New York: John Wiley, pp. 3-20.

Chi, E.M. et Reinsel, G.C. (1989), "Models for Longitudinal Data with Random Effects and AR(1) Errors," *Journal of the American Statistical Association*, 84, 452-459.

Choudhry, H. et Hunter, L. (1987), "Modélisation de séries chronologiques pour l'établissement d'estimations régionales," dans *Les utilisations statistiques des données administratives*: Recueil, Ottawa, Statistique Canada.

7. ÉTUDE EMPIRIQUE

Nous évaluons maintenant les efficacités du MPLNBE ainsi que celles des deux estimateurs synthétiques et de l'estimateur d'enquête y_{it}, à l'aide des données pour une période de 36 mois (janvier 1983 - décembre 1985) d'estimations d'enquête du chômage tirées de l'enquête sur la population active du Canada pour 21 divisions de recensement (petites régions) dans la province de la Colombie-Britannique. Les variables auxiliaires utilisées dans la régression sont des données administratives mensuelles provenant du système d'AC et le nombre de personnes dans la population active tiré des données de l'enquête sur la population active. Posant ici, $t = 1, \ldots, 36$ et $i = 1, \ldots, 21$, y_{it} = log (estimation de la proportion de la population en chômage d'après l'enquête), x_{1it} = log (bénéficiaires de l'assurance-chômage/nombre projeté de personnes de 15 ans et plus), x_{2it} = estimation du taux d'activité d'après l'enquête. Le taux d'activité est défini comme la proportion de la population cible qui est soit occupée, soit en chômage. Bien que x_{2it} soit sujette à des erreurs d'échantillonnage, son coefficient de variation (cv) est négligeable si on le compare à celui de y_{it}, on peut donc ne pas tenir compte de ces erreurs sans que cela ait un effet sur les estimations.

Notre modèle (3.9), avec deux variables concomitantes peut être écrit de la façon suivante

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + v_i + w_{it}$$

$$w_{it} = \rho w_{i,t-1} + \epsilon_{it}, \ |\rho| < 1 \tag{7.1}$$

où $v_i \sim_{ind} N(0,\sigma_v^2)$ et $\epsilon_{it} \sim_{ind} N(0,\sigma^2)$. L'EQM estimée du MPLNBE selon (7.1) a été calculée à partir de (5.1) pour chaque (i, t) en y introduisant les estimations $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})$ et en utilisant $(1, x_{1it}, x_{2it})$ pour la t^e ligne de X_i. Ces estimateurs ont été obtenus au moyen de la méthode de Pantula et Pollock (1985), et en voici les valeurs:

$$\hat{\sigma}^2 = 0.0391, \quad \hat{\sigma}_v^2 = 0.0175, \quad \hat{\rho} = 0.362.$$

Passant à l'estimateur synthétique, son EQM estimée, $\hat{\theta}_{it}(S)$, qui ne tient pas compte des effets aléatoires $\{v_i\}$ dans (7.1), est obtenue pour chaque (i, t) au moyen de (6.3) en y introduisant $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})$. De même, l'EQM estimée de l'estimateur synthétique $\hat{\theta}_{it}(S1)$, qui traite $\{v_i\}$ comme des effets fixes dans (7.1) est obtenue pour chaque (i, t) à partir de (6.6) en remplaçant β par $\hat{\beta}$ puis en utilisant $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})$ à la place de $(\sigma^2, \sigma_v^2, \rho)$. Finalement, on obtient une estimation de l'EQM de l'estimation d'enquête de y_{it} au moyen de (5.2) en utilisant $(\hat{\sigma}^2, \hat{\rho})$ à la place de (σ^2, ρ).

Représentons l'efficacité estimée du MPLNBE par rapport à l'estimateur synthétique $\hat{\theta}_{it}(S)$ par E_{1it} = EQM est $[\hat{\theta}_{it}(S)]$/EQM est (MPLNBE), l'efficacité estimée du MPLNBE par rapport à l'estimateur synthétique $\hat{\theta}_{it}(S1)$ par E_{2it} = EQM est MSE$[\hat{\theta}_{it}(S1)]$/EQM est (MPLNBE) et l'efficacité du MPLNBE par rapport à l'estimation d'enquête y_{it} par E_{3it} = EQM est(y_{it})/EQM est (MPLNBE). Il faut remarquer que les EQM des estimateurs synthétiques $\hat{\theta}_{it}(S)$ et $\hat{\theta}_{it}(S1)$ et de l'estimation d'enquête y_{it} sont calculées selon le modèle (7.1).

Les moyennes de E_{1it}, de E_{2it} de E_{3it} pour trente-six mois sont calculées comme suit: $\bar{E}_{1i} = \sum_t E_{1it}/36$, $\bar{E}_{2i} = \sum_t E_{2it}/36$, et $\bar{E}_{3i} = \sum_t E_{3it}/36$ pour chaque petite région i ces valeurs sont données dans le tableau 1.

Il est évident, quand on consulte le tableau 1, que le MPLNBE permet d'accroître considérablement l'efficacité moyenne par rapport à l'estimateur d'enquête, $\bar{E}_{3i}$ étant compris entre 7.56 et 10.11. Les augmentations de l'efficacité moyenne du MPLNBE par rapport à l'estimateur synthétique $\hat{\theta}_{it}(S)$ sont aussi substantielles, $\bar{E}_{1i}$ allant de 2.67 à 3.56. L'efficacité moyenne du MPLNBE par rapport à l'estimateur synthétique $\hat{\theta}_{it}(S1)$ représentée par $\bar{E}_{2i}$ est comprise entre 0.82 et 6.45. Voici les valeurs moyennes globales de l'efficacité: $\bar{E}_1 = \sum \bar{E}_{1i}/21 = 3.02$, $\bar{E}_2 = \sum \bar{E}_{2i}/21 = 2.50$ et $\bar{E}_3 = \sum \bar{E}_{3i}/21 = 8.61$.

$$\tilde{\theta}_{it}(S) = \tilde{\beta}_0(S) + \tilde{\beta}_1(S)x_{it}, \tag{6.2}$$

où $\tilde{\beta}_0(S)$et $\tilde{\beta}_1(S)$ sont les estimateurs des moindres carrés généralisés de β_0 et de β_1 selon le modèle (6.1):

$\tilde{\beta}(S) = (X'R^{*-1}X)^{-1})(X'R^{*-1}y)$, où R^* est donnée par R où ρ^* remplace ρ. L'estimateur $\tilde{\theta}_{it}(S)$ est non biaisé pour $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i$ selon le modèle qui nous intéresse, (3.9).

Si nous écrivons $\tilde{\theta}_{it}(s)$ comme une fonction linéaire, $\mathbf{a}'\mathbf{y}$, des observations $\mathbf{y}$, on peut obtenir l'EQM de $\tilde{\theta}_{it}(S)$ selon le modèle qui nous intéresse. Cette erreur est donnée par

$$\begin{aligned} \mathrm{MSE}[\tilde{\theta}_{it}(S)] &= E(\mathbf{a}'\mathbf{y} - \mathbf{k}'\beta - \mathbf{m}'\mathbf{v})^2 \\ &= \sigma^2[(Z'\mathbf{a} - \mathbf{m})'(Z'\mathbf{a} - \mathbf{m})(\sigma_v^2/\sigma^2) + \mathbf{a}'R\mathbf{a}], \end{aligned} \tag{6.3}$$

où $\mathbf{k}$ et $\mathbf{m}$ sont donnés par (4.3).

Puisque $\tilde{\theta}_{it}(S)$ dépend des coefficients d'autocorrélation inconnus ρ^*, nous estimons ρ^* à l'aide de (6.1) au moyen de la méthode de Gauss-Newton modifiée (Hartley, 1961). L'EQM de l'estimateur résultant, $\tilde{\theta}_{it}(S)$, est estimée en utilisant $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})(\hat{\sigma}_v^2, \hat{\rho})$ à la place de dans (6.3). Cet estimateur sous-estimera la vraie EQM de $\hat{\theta}_{it}(S)$ puisqu'on ne tient pas compte de l'incertitude reliée à l'estimation de ρ^*. Néanmoins, la sous-estimation ne sera vraisemblablement pas grave pour notre étude empirique décrite dans la section 7.

On obtient un autre estimateur synthétique en considérant le modèle d'effets fixes (3.2) pour θ_{it}, puis en écrivant

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + v_i + \tilde{w}_{it} \\ \tilde{w}_{it} &= \tilde{\rho}\tilde{w}_{i,t-1} + \tilde{\varepsilon}_{it} \; |\tilde{\rho}| < 1, \end{aligned} \tag{6.4}$$

où $\tilde{\varepsilon}_{it} \sim_{\text{ind}} N(0, \tilde{\sigma}^2)$ et $\{v_i\}$ sont des effets fixes pour petites régions. L'estimateur synthétique de θ_{it} résultant est donné par (Choudhry et Hunter, 1987):

$$\tilde{\theta}_{it}(S1) = \tilde{\beta}_0(S1) + \tilde{\beta}_1(S1)x_{it} + \tilde{v}_i(S1), \tag{6.5}$$

où $\tilde{\gamma}(S1) = [\tilde{\beta}_0(S1), \tilde{\beta}_1(S1), \tilde{v}_1(S1), \ldots, (S1)]'$ est l'estimateur des moindres carrés généralisé donné par $(\mathbf{W}'\tilde{R}^{-1}\mathbf{W})^-$ $(\mathbf{W}'\tilde{R}^{-1}y)$. Ici la $(i, t)^e$ ligne de $\mathbf{W}$ est le vecteur $1 \times (I + 2)$ $(1, x_{it}, 0, \ldots, 0, 1, 0, \ldots, 0)$ avec 1 dans la $(i + 2)^e$ position, $\tilde{R}$ est donnée par R où $\tilde{\rho}$ remplace ρ et $(\mathbf{W}'\tilde{R}^{-1}\mathbf{W})^-$ est une matrice pseudo-inverse de $\mathbf{W}'\tilde{R}^{-1}W$. L'estimateur $\theta_{it}(S1)$ est unique pour tout choix de matrice pseudo-inverse.

En écrivant $\tilde{\theta}_{it}(S1)$ comme $\mathbf{b}'y$, on voit que $\tilde{\theta}_{it}(S1)$ est biaisée pour θ_{it} selon le modèle qui nous intéresse, (3.9). Dans ce cas, son EQM est donnée par

$$\begin{aligned} \mathrm{MSE}[\tilde{\theta}_{it}(S1)] &= E(\mathbf{b}'y - \mathbf{k}'\beta - \mathbf{m}'v)^2 \\ &= [(X'\mathbf{b} - \mathbf{k})'\beta]^2 + \sigma^2[(Z'\mathbf{b} - \mathbf{m})'(Z'\mathbf{b} - \mathbf{m})(\sigma_v^2/\sigma^2) + \mathbf{b}'R\mathbf{b}] \end{aligned} \tag{6.6}$$

Puisque l'estimateur $\tilde{\theta}_{it}(S1)$ dépend du coefficient d'autocorrélation inconnu $\tilde{\rho}$, nous estimons $\tilde{\rho}$ à partir de (6.4) à l'aide de la méthode de Gauss-Newton modifiée. $\tilde{\theta}_{it}(s1)$ L'EQM de l'estimateur résultant, $\tilde{\theta}_{it}(s1)$, est estimée en remplaçant β par $\tilde{\beta}$ dans (6.6) puis en utilisant $(\hat{\sigma}^2, \hat{\sigma}_v^2, \hat{\rho})$ à la place de $(\sigma^2, \sigma_v^2, \rho)$. Cet estimateur sous-estimera la vraie EQM de $\hat{\theta}_{it}(S1)$ puisqu'on ne tient pas compte de l'incertitude reliée à l'estimation de $\tilde{\rho}$. Néanmoins, la sous-estimation ne sera vraisemblablement pas grave pour notre étude empirique décrite dans la section 7.

Si p-1($\geq$2) variables auxiliaires sont incluses dans le modèle, alors on obtient les $(\tilde{e}_{it})$ en faisant la régression des y_{it} par rapport aux $x_{1it}, \ldots, x_{p-1,it}$, y compris l'ordonnée à l'origine. De même, $(h_{jit}, h_{jit}^{(1)}, h_{jit}^{(2)})$, $j = 0,1, \ldots, p-1$ sont définis en fonction des éléments $1, x_{1it}, \ldots, x_{p-1,it}$, et $\hat{e}'\hat{e}$ est obtenue en faisant la régression de $z_{it}^{(1)}$ par rapport à $h_{0it}^{(1)}, h_{1it}^{(1)}, \ldots, h_{p-1,it}^{(1)}$ sans l'ordonnée à l'origine, et $\hat{u}'\hat{u}$ est obtenue en définissant $f_{ji} = \sum_{t=1}^{T} f_t h_{jit}$, $j = 0,1, \ldots, p-1$ et en faisant la régression de g_i par rapport à $f_{0i}, f_{1i}, \ldots, f_{p-1,i}$ sans l'ordonnée à l'origine. Finalement, $\hat{\sigma}^2$ et $\hat{\sigma}_v^2$ sont définis par (4.6) et (4.7) respectivement, avec $I(T-1) - 2$ remplacé par $I(T-1) - p$ et $I - 2$ remplacé par $I - p$.. Il est aussi possible d'obtenir les estimations du maximum de vraisemblance de σ^2, σ_v^2 et ρ, à l'aide de l'algorithme EM (voir Chi et Reinsel, 1989).

Si nous introduisons les estimateurs $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ et $\hat{\rho}$ dans (4.2) à la place des variables correspondantes, nous obtenons le meilleur prédicteur linéaire non biaisé empirique (MPLNBE) de θ_{it}, représenté par $\hat{\theta}_{it}$.

5. ERREUR QUADRATIQUE MOYENNE DU MPLNBE

Suivant Henderson (1975), l'erreur quadratique moyenne (EQM) du MPLNB, $\tilde{\tau} = \tilde{\theta}_{it}$, , est donnée par

$$\text{EQM } (\tilde{\theta}_{it}) = \sigma^2 \{k'(X'\Sigma^{-1}X)^{-1} k + (\sigma_v^2/\sigma^2)m'm - (\theta_v^2/\theta^2)^2 \; m'Z'\Sigma^{-1} AZm$$

$$- 2(\sigma_v^2/\sigma^2) \; k'(X'\Sigma^{-1}X)^{-1} \; X'\Sigma^{-1}Zm\}, \tag{5.1}$$

où $A = I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$. L'EQM du MPLNBE, $\hat{\theta}_{it}$, comprend des termes d'ordre inférieur qui tiennent compte de l'incertitude dans les estimateurs $\hat{\theta}^2$, $\hat{\theta}_v^2$ et $\hat{\rho}$. Nous sommes à élaborer une approximation précise de l'EQM du MPLNBE d'après la méthode proposée par Prasad et Rao (1990) pour le modèle de Fay-Herriot.

Dans le présent article, nous n'avons pas tenu compte de l'incertitude dans les estimateurs $\hat{\theta}^2$, $\hat{\theta}_v^2$ et $\hat{\rho}$, et nous avons utilisé (5.1) avec $(\hat{\theta}^2, \hat{\theta}_v^2 \; \hat{\rho})$ à la place de $(\theta^2, \theta_v^2, \rho)$ comme estimateur de l'EQM du MPLNBE. Cet estimateur sous-estime la véritable EQM du MPLNBE, mais la sous-estimation ne sera vraisemblablement pas grave pour notre étude empirique décrite dans la section 7.

L'EQM de l'estimateur d'enquête, y_{it}, de θ_{it}, selon le modèle (3.9), est donnée par

$$\text{EQM } (y_{it}) = E \, (y_{it} - \theta_{it})^2 = V(w_{it}) = \theta^2/(1 - \rho^2) \tag{5.2}$$

Un estimateur de l'EQM (y_{it}) est obtenu en utilisant $(\hat{\theta}^2, \hat{\rho})$ à la place de (θ^2, ρ) dans (5.2).

6. ESTIMATEURS SYNTHÉTIQUES

Si, en plus de ne pas tenir compte des effets aléatoires pour petites régions (v_i), nous utilisons le modèle

$$y_{it} = \beta_0 + \beta_1 x_{it} + w_{it}^{\star},$$

$$w_{it}^{\star} = \rho^{\star} w_{i,t-1}^{\star} + \varepsilon_{it}^{\star}, \; |\rho^{\star}| < 1, \tag{6.1}$$

où $\varepsilon_{it}^{\star} \sim_{ind} N(0, \theta^{\star 2})$, nous obtenons un estimateur synthétique de $\theta_{it} = \beta_0 + \beta_1 x_{it}$. Cet estimateur est donné par

4.2 Estimation de σ^2

Le MPLNB (4.2) dépend du rapport des variances σ_v^2/σ^2 inconnu et du coefficient d'autocorrélation inconnu ρ. Nous avons utilisé la méthode de Pantula et Pollack (1985) pour estimer les paramètres σ^2, σ_v^2 et ρ. Cette méthode constitue un prolongement de la méthode d'ajustement des constantes pour le cas $\rho = 0$ (Fuller et Battese, 1973) et les estimations de σ^2, σ_v^2 et ρ sont obtenues à l'aide des formules données plus loin.

Soit Lt $(\tilde{e}_{it})$ les résidus des moindres carrés ordinaires obtenus en faisant la régression des y_{it} par rapport aux x_{it}, y compris l'ordonnée à l'origine. Alors ρ est estimé par

$$\hat{\rho} = \left[\sum_{i=1}^{I}\sum_{t=1}^{T-2} \tilde{e}_{it}(\tilde{e}_{i,t+1} - \tilde{e}_{i,t+2})\right]\left[\sum_{i=1}^{I}\sum_{t=1}^{T-2} \tilde{e}_{it}\,(\tilde{e}_{it} - \tilde{e}_{i,t+1})\right]^{-1}. \tag{4.5}$$

Définissons

$$z_{it}^{(1)} = z_{it} - z_{it}^{(2)},$$

où

$$z_{it} = y_{it} - \hat{\rho}y_{i,t-1},\ t \geq 2$$
$$= f_1 y_{it},\ t = 1$$

et

$$z_{it}^{(2)} = c^{-1}d_i f_t$$

avec

$$c = (1 - \hat{\rho})[T - (T - 2)\hat{\rho}],$$

$$f_t = 1 - \hat{\rho}^2,\ t = 1$$
$$= 1 - \hat{\rho},\ \ t \geq 2$$

et

$$d_i = \sum_{t=1}^{T} f_t z_{it}.$$

De même, définissons $(h_{0it},\ h_{0it}^{(1)},\ h_{0it}^{(2)})$ et $h_{1it},\ h_{1it}^{(1)},\ h_{1it}^{(2)})$ en fonction des éléments 1 et x_{it}, c.-à-d. remplaçons y_{it} par 1 et x_{it} respectivement dans les expressions pour $(z_{it},\ z_{it}^{(1)},\ z_{it}^{(2)})$. Soit $\hat{e}'\hat{e}$ la somme des carrés des résidus obtenue en faisant la régression de $z_{it}^{(1)}$ par rapport à $h_{0it}^{(1)}$ et à $h_{1it}^{(1)}$, sans l'ordonnée à l'origine. Définissons aussi

$$g_i = \sum_{t=1}^{T} f_t z_{it},$$

$$f_{0i} = \sum_{t=1}^{T} f_t h_{0it},\ f_{1i} = \sum_{t=1}^{T} f_t h_{1it}.$$

Soit $\hat{u}'\hat{u}$, la somme des carrés des résidus obtenue en faisant la régression de g_i par rapport à f_{0i} et à f_{1i}, sans l'ordonnée à l'origine. Les estimations de σ^2 et de σ_v^2 sont maintenant obtenues à l'aide des formules suivantes

$$\sigma^2 = [I(t - 1) - 2]^{-1}\hat{e}'\hat{e} \tag{4.6}$$

et

$$\hat{\sigma}_v^2 = c^{-1}(I - 2)^{-1}[\hat{u}'\hat{u} - \hat{\sigma}^2(I - 2)], \tag{4.7}$$

dans le cas du modèle (3.9).

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + (e_{it} + u_{it}),$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1, \tag{3.7}$$

où $v_i \sim_{ind} N(0,\sigma_v^2)$, $\varepsilon_{it} \sim_{ind} N(0,\sigma^2)$, et les e_{it} ont une moyenne zéro et une matrice des covariances diagonale par blocs $\sum = \text{diag}(\sum_1, \ldots, \sum_I)$ connue.

Malheureusement, la matrice des covariances d'échantillonnage $\sum$ provenant de l'enquête sur la population active du Canada n'est pas disponible actuellement, de sorte que nous avons traité l'erreur composite $w_{it} = e_{it} + u_{it}$ comme un processus AR(1): où $w_{it} = \rho w_{i,t-1} + \varepsilon_{it}$ avec $\varepsilon_{it} \sim_{ind} N(0,\sigma^2)$ et nous avons alors considéré θ_{it} comme

$$\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i. \tag{3.8}$$

Tiller (1989) a utilisé une approche semblable dans le contexte de l'estimation de la population active à partir de données chronologiques agrégées qui découlent d'enquêtes répétées. On peut écrire le modèle combiné de la façon mentionnée ci-dessous, suivant les hypothèses qui précèdent

$$y_{it} = \beta_0 + \beta_0 x_{it} + v_i + w_{it},$$

$$w_{it} = \rho w_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1, \tag{3.9}$$

où $v_i \sim_{ind} N(0,\sigma_v^2)$ et $\varepsilon_{it} \sim_{ind} N(0,\sigma^2)$.

4. MEILLEUR PRÉDICTEUR LINÉAIRE NON BIAISÉ EMPIRIQUE

4.1 MPLNB

Si l'on écrit les données $\{y_{it}\}$ sous la forme $y = (y_{11}, \ldots, y_{1T}; \ldots; y_{I1}, \ldots, y_{IT})' = (y_1', \ldots, y_I')'$, le modèle (3.9) peut être exprimé comme un cas spécial du modèle mixte général

$$y = X\beta + Zv + w \tag{4.1}$$

avec

$$X' = (X_1', \ldots, X_I')$$

$$Z = I \otimes 1_T, \quad \beta = (\beta_0, \beta_1)',$$

où X_i est une matrice $T \times 2$ dont la t^e ligne est donnée par $(1, x_{it})$, I est la matrice unité d'ordre I et 1_T est le t-vecteur de 1. De plus,

$$E(v) = 0, \quad \text{Cov}(v) = \sigma_v^2 I$$

$$E(w) = 0, \quad \text{Cov}(w) = \sigma^2(I \otimes \Gamma) = \sigma^2 R \text{ (say)}$$

et Γ est une matrice $T \times T$ avec élément (i,j) $\gamma_{ij} = (1 - \rho^2)^{-1} \rho^{|i-j|}$.

En calculant le meilleur prédicteur linéaire non biaisé (MPLNB) de toute combinaison linéaire de β et des effets aléatoires v, disons $\tau = k'\beta + m'v$ Henderson (1975) obtient

$$\tilde{\tau} = k'\tilde{\beta} + m'Z'\textstyle\sum^{-1}(y - X\tilde{\beta})(\sigma_v^2/\sigma^2). \tag{4.2}$$

Ici, $\sum = I \otimes [(\sigma_v^2/\sigma^2) J + \Gamma]$ où J représente une matrice $T \times T$ de 1 et $\hat{\beta} = (X'\sum^{-1}X)^{-1}(X'\sum^{-1}y)$ est l'estimateur des moindres carrés généralisés de β. Si $\tau = \theta_{it}$ comme (3.8) le donne, alors

$$k' = (1, x_{it}), \quad m' = (0, \ldots, 0, 1, 0, \ldots, 0) \tag{4.3}$$

avec 1 dans la i^e position et

$$m'Z'\textstyle\sum^{-1}(y - X\tilde{\beta}) = 1_T'[(\sigma_v^2/\sigma^2)J + \Gamma]^{-1}(y_i - X_i\tilde{\beta}). \tag{4.4}$$

qu'un Etat, à l'aide d'estimations d'enquête mensuelles de la Current Population Survey (CPS) (enquête sur la population actuelle) comme variables dépendantes et de données tirées du système d'AC ainsi que de variables du recensement comme variables indépendantes. Notre recherche visait à obtenir des estimations mensuelles fiables du chômage pour les divisions de recensement, à l'aide des estimations des taux de chômage et d'activité obtenues dans le cadre de l'enquête sur la population active et de données administratives tirées du système de l'AC. On utilise les taux moyens de chômage pour trois ans pour les divisions de recensement conjointement avec d'autres variables afin de produire un indice qui, à son tour, est employé pour affecter des fonds à des fins de subventions à l'industrie.

Il existe un nombre considérable de documents relatifs à l'économétrie qui portent sur la modélisation et l'estimation des liens qui combinent des données chronologiques et des données transversales (par exemple, voir Judge et coll. 1980, chapitre 13), mais l'on tient rarement compte des erreurs d'échantillonnage. Nous allons maintenant considérer certains de ces modèles. Pour simplifier, nous ne considérons, encore une fois, qu'une variable concomitante. Soit θ_{it}, y_{it} et x_{it} qui représentent respectivement la moyenne de la population, l'estimateur d'enquête direct et la variable concomitante associées à la i^e petite région à la période $t(i = 1, \ldots, I;\ t = 1, \ldots, T)$. Nous avons

$$y_{it} = \theta_{it} + e_{it},\ i = 1, \ldots I;\ t = 1, \ldots, T, \qquad (3.1)$$

et, suivant Fay et Herriot (1979), nous supposons que la matrice des covariances des erreurs d'échantillonnage e_{it} est une matrice diagonale par blocs avec des blocs connus Σ_i, où Σ_i est une matrice $T \times T$ et $E(e_{it}) = 0$. Les recherches récentes se sont concentrées sur la modélisation des erreurs d'échantillonnage des données agrégées. Par exemple, Binder et Dick (1989) et Tiller (1989) ont proposé des modèles autorégressifs à moyennes mobiles (ARMM).

Les modèles pour θ_{it}, proposés dans les documents sur l'économétrie, comprennent ceux qui suivent:

$$\text{(I)}\quad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \varepsilon_{it}, \qquad (3.2)$$

où les v_i sont des effets fixes pour petites régions et les ε_{it} sont des variables normales indépendantes avec une moyenne nulle et une variance σ^2, ce qui est représenté sous forme abrégée par $\varepsilon_{it} \backsim_{\text{ind}} N(0,\sigma^2)$.

$$\text{(II)}\quad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + \varepsilon_{it}, \qquad (3.3)$$

où $v_i \backsim_{\text{ind}} N(0,\sigma_v^2)$, $\varepsilon_{it} \backsim_{\text{ind}} N(0,\sigma^2)$ et $\{v_i\}$ et $\{\varepsilon_{it}\}$ sont indépendants. Ici les v_i sont des effets aléatoires pour petites régions.

$$\text{(III)}\quad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_t + \varepsilon_{it}, \qquad (3.4)$$

où $v_i \backsim_{\text{ind}} N(0,\sigma_v^2)$, $u_t \backsim_{\text{ind}} N(0,\sigma_u^2)$, $\varepsilon_{it} \backsim_{\text{ind}} N(0,\sigma^2)$ et $\{v_i\}$, $\{u_t\}$ $\{\varepsilon_{it}\}$ sont indépendants. Ici les v_i et les u_i sont des effets aléatoires pour petites régions et des effets temporels aléatoires respectivement.

$$\text{(IV)}\quad \theta_{it} = \beta_0 + \beta_1 x_{it} + v_i + u_{it}, \qquad (3.5)$$

et

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it},\ |\rho| < 1$$

où $v_i \backsim_{\text{ind}} N(0,\sigma_v^2)$, $\varepsilon_{it} \backsim_{\text{ind}} N(0,\sigma^2)$ et $\{v_i\}$, $\{\varepsilon_{it}\}$ sont indépendants. Ici les v_i sont des effets aléatoires pour petites régions et les $\{u_{it}\}$ suivent un processus AR(1). Le modèle (3.5) peut être réécrit sous forme d'un modèle à retard échelonné:

$$\theta_{it} = \rho\theta_{i,t-1} + (1 - \rho)\beta_0 + \beta_1 x_{it} - \beta_1 \rho x_{i,t-1} + (1 - \rho)v_i + \varepsilon_{it}. \qquad (3.6)$$

Le modèle IV semble être le plus réaliste des quatre modèles étudiés puisque l'autre forme qu'il peut prendre (3.6) établit un lien entre la moyenne de la population courante, θ_{it}, et la moyenne de la population pour la période précédente, $\theta_{i,t-1}$, ainsi que les valeurs de la variable auxiliaire pour les périodes courante et précédente, x_{it} et $x_{i,t-1}$ respectivement. La forme (3.5) du modèle IV reflète la dépendance de θ_{it} dans le temps pour chaque région i. Dorénavant, nous adoptons le modèle IV sous la forme (3.5).

Le modèle combiné est donné par les équations ci-après, si nous utilisons (3.1) et (3.5)

Nous supposons que le modèle de régression linéaire ci-après pour θ_i lie les petites régions au moyen des données concomitantes x_i:

$$\theta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, I, \tag{2.1}$$

où β_0 et β_1 sont les coefficients de régression. Un estimateur par régressions synthétique de θ_i est alors donné par

$$\tilde{\theta}_{i(reg)} = \tilde{\beta}_0 + \tilde{\beta}_1 x_i, \tag{2.2}$$

où $\tilde{\beta}_0$ et $\tilde{\beta}_1$ sont les estimateurs des moindres carrés ordinaires de β_0 et de β_1 obtenus à partir du modèle combiné $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \ldots, I$. Nous pouvons aussi utiliser les estimateurs des moindres carrés généralisés (pondérés) de β_0 et β_1 si la matrice des covariances estimées des estimateurs d'enquête y_i est disponible.

L'estimateur synthétique (2.2) pourrait entraîner des biais considérables puisqu'il ne donne pas de poids à l'estimateur d'enquête direct y_i. Par contre, l'estimateur empirique de Bayes ou le MPLNBE donne les poids appropriés à l'estimateur d'enquête ainsi qu'à l'estimateur synthétique et, par conséquent, entraîne des biais plus petits par rapport à l'estimateur synthétique.

2.2 Estimateur empirique de Bayes ou MPLNBE

Fay et Herriot (1979) ont introduit l'incertitude dans le modèle 2.1 de la façon suivante:

$$\theta_i = \beta_0 + \beta_1 x_i + v_i, \tag{2.3}$$

où les v_i sont des variables normales indépendantes avec une moyenne nulle et une variance inconnue σ_v^2. Pour les erreurs d'échantillonnage, ils ont supposé que les e_i sont des variables normales indépendantes avec $E(e_i) = 0$ et $Var(e_i) = \sigma_i^2$, où σ_i^2 est connue. Le modèle combiné est donné par

$$y_i = \beta_0 + \beta_1 x_i + v_i + e_i. \tag{2.4}$$

L'estimateur empirique de Bayes de θ_i est donné sous forme d'une somme pondérée de l'estimateur d'enquête direct y_i et de l'estimateur par régression synthétique $\hat{\theta}_{i(reg)} = \hat{\beta}_0 + \hat{\beta}_1 x_1$:

$$t_i(\hat{\sigma}_v^2, y) = w_i y_i + (1 - w_i)\hat{\theta}_i(\text{reg}), \tag{2.5}$$

où $w_i = \hat{\sigma}_v^2/\hat{\sigma}_v^2 + \sigma_i^2)$ et $\hat{\beta}_0$ ainsi que $\hat{\beta}_1$ sont les estimateurs des moindres carrés pondérés selon le modèle combiné et $\hat{\sigma}_v^2$ est un estimateur de σ_v^2. On peut utiliser un estimateur de moment simple de σ_v^2 ou un estimateur plus complexe, comme l'estimateur du maximum de vraisemblance de σ_v^2. Fay et Herriot (1979) ont utilisé (2.5) pour estimer le revenu par habitant pour de petites régions (c.-à-d. pour les populations composées de moins de 1,000 personnes) à partir des données du U.S. Census of Population and Housing (recensement de la population et du logement des Etats-Unis) de 1970 et ils ont présenté des preuves que (2.5) entraîne une erreur moyenne plus petite que soit l'estimateur d'enquête direct, soit l'estimateur synthétique qui utilise la moyenne pour le comté.

Prasad et Rao (1990) ont obtenu un estimateur précis de l'erreur quadratique moyenne du MPLNBE (2.5) en tenant compte de l'incertitude dans l'estimateur de σ_v^2.

3. MODÈLES TRANSVERSAUX ET DE SÉRIES CHRONOLOGIQUES

Les méthodes décrites dans la section 2 n'utilisent que des données transversales à un moment donné et, par conséquent, elles n'exploitent pas les renseignements que renferment les données à d'autres moments. Scott et coll. (1977), Jones (1980), Tiller (1989) et d'autres ont utilisé la modélisation de séries chronologiques de données agrégées (p. ex., des moyennes globales) à partir de données d'enquêtes répétées et ont obtenu des estimateurs améliorés des données agrégées à différents moments. Cependant, il existe très peu d'articles portant sur des travaux visant à étendre la méthode de Fay-Herriot pour l'estimation relative à de petites régions à des séries chronologiques d'estimations d'enquête transversales portant sur de petites régions conjointement avec des données du recensement et des données supplémentaires, qui varient dans le temps, comme les donnée tirées de dossiers administratifs.

Cronkhite (1986) a élaboré des estimateurs par régression synthétiques à l'aide de données chronologiques transversales combinées et il les a appliqués afin d'estimer l'emploi et le chômage dans des régions plus petites

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ESTIMATION DE DONNÉES RÉGIONALES À L'AIDE DE MODÈLES QUI COMBINENT DES SÉRIES CHRONOLOGIQUES ET DES DONNÉES TRANSVERSALES

G.H. Choudhry[1] et J.N.K. Rao[2]

RÉSUMÉ

On élabore des modèles transversaux et de séries chronologiques avec effets aléatoires et erreurs autocorrélées. A l'aide de ces modèles, on obtient les "meilleurs estimateurs linéaires non biaisés" pour de petites régions à chaque moment. Les efficacités de plusieurs estimateurs pour de petites régions sont évaluées à l'aide d'estimations du chômage obtenues au moyen d'enquêtes mensuelles pour des divisions de recensement (petites régions) provenant de l'enquête sur la population active du Canada combinées avec des données administratives mensuelles tirées du système de l'assurance-chômage et des estimations de la population active obtenues à l'aide d'enquêtes mensuelles, comme variables auxiliaires.

1. INTRODUCTION

La demande pour des données régionales fiables a augmenté régulièrement au cours des dernières années parce que ces données sont utilisées pour formuler des politiques et des programmes, pour affecter les fonds gouvernementaux et dans des programmes régionaux. Pour répondre aux besoins des utilisateurs, Statistique Canada a entrepris un programme d'élaboration en rapport avec les petites régions. Brackstone (1986) a discuté des questions qui découlent de l'élaboration et de la fourniture de données pour les petites régions.

Les estimateurs régionaux directs tirés de données d'enquête donneront vraisemblablement lieu à des erreurs-types inacceptables, parce que trop grosses, à cause des petites tailles des échantillons. On a donc besoin d'autres estimateurs qui "empruntent" à de petites régions connexes afin d'améliorer l'efficacité. De tels estimateurs utilisent, soit implicitement, soit explicitement, des modèles qui relient les petites régions à l'aide de données supplémentaires comme des données du dernier recensement et d'autres tirées de dossiers administratifs.

La majorité des recherches relatives à l'estimation pour de petites régions se sont concentrées sur les données transversales à un moment donné. Rao (1986) a fait un exposé de ces recherches. Les estimateurs proposés dans les documents portant sur le sujet comprennent a) des estimateurs synthétiques (Gonzalez, 1973; Ericksen, 1974), des estimateurs qui préservent la structure (EQPS --"SPREE"), Purcell et Kish (1980); b) des estimateurs dépendant de la taille de l'échantillon (Drew et coll. 1982; Särndal et Hidiroglou,1989); c) des estimateurs empiriques de Bayes (Fay et Herriot, 1979) et des meilleurs prédicteurs linéaires non biaisés empiriques (MPLNBE), Battesse et coll. (1988) et Prasad et Rao (1990). On obtient le MPLNBE à partir du meilleur prédicteur linéaire non biaisé (MPLNB) en remplaçant les paramètres de la variance inconnus par leurs estimateurs, comme on obtient l'estimateur empirique de Bayes à partir de l'estimateur de Bayes.

Le but principal du présent article est d'élaborer des modèles transversaux et de séries chronologiques avec effets aléatoires et erreurs autocorrélées et d'obtenir des MPLNBE pour de petites régions à chaque moment, à l'aide de ces modèles. Dans la section 2 nous passons en revue le travail effectué sur les estimateurs synthétiques par régression et sur les estimateurs empiriques de Bayes obtenus à partir de données transversales à un moment donné. Les modèles transversaux et de séries chronologiques sont étudiés dans la section 3 et on propose un prolongement du modèle de Fay-Herriot (1979). Le MPLNBE est obtenu dans la section 4. Les efficacités du MPLNBE, par rapport à deux estimateurs synthétiques et à un estimateur d'enquête direct, sont évaluées dans la section 5, à l'aide d'estimations d'enquête mensuelles du chômage pour des divisions de recensement (petites régions) provenant de l'enquête sur la population active du Canada combinées avec des données administratives mensuelles provenant du système de l'assurance-chômage (AC) et des estimations d'enquête mensuelles de la population active comme variables auxiliaires.

2. ESTIMATEURS TRANSVERSAUX

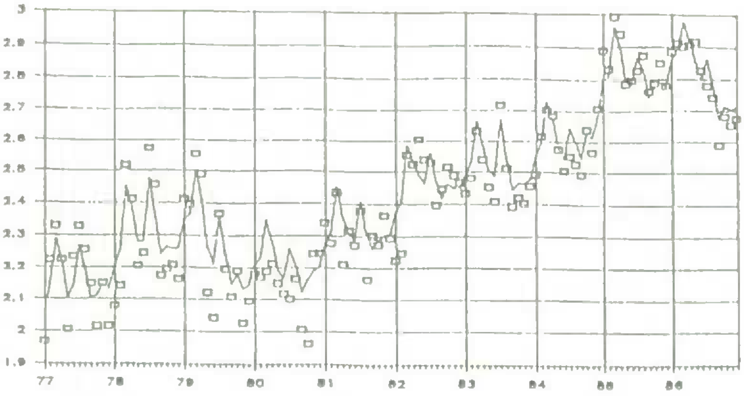
2.1 Estimateurs synthétiques par régression

Soit y_i l'estimateur d'enquête direct de la moyenne θ_i de la i^e petite région à un moment donné. Pour simplifier, nous supposons qu'une seule variable concomitante x_i apparentée à θ_i est disponible; le prolongement à deux variables concomitantes ou plus est simple. Nous supposons aussi que y_i est non biaisé pour θ_i, c.-à-d. que $y_i = \theta_i + e_i$ où les e_i sont les erreurs d'échantillonnage avec $E(e_i) = 0$.

[1] G.H. Choudhry, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6
[2] J.N.K. Rao, Département de mathématiques et de statistiques, Carleton University, Ottawa (Ontario) K1S 5B6

Graphique 2a

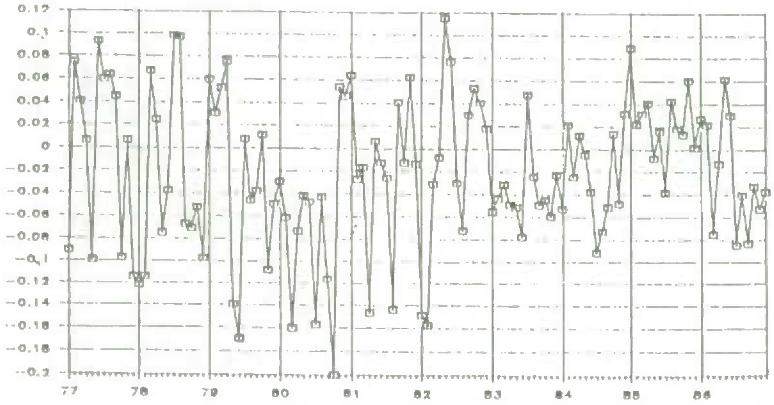
ICB: Chômage 1977-1986



Année
Observé (log)

Graphique 2b

ICB: Chômage 1977-1986
Valeurs observées - Lissées



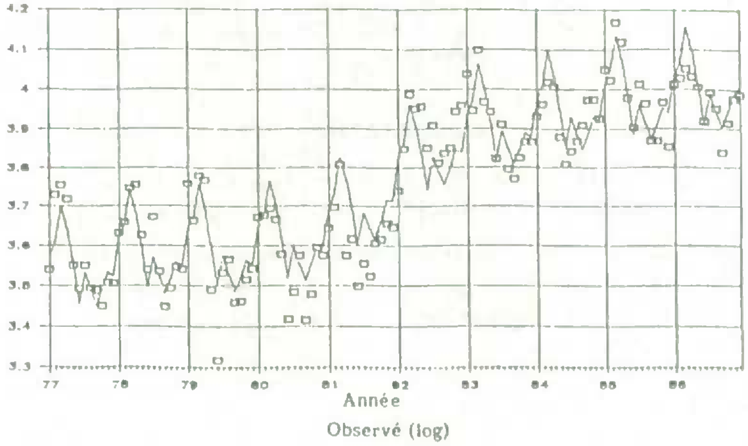
Année

Tableau 1: Estimations des paramètres - Séries de chômage 1977-1988

	Nouvelle Écosse				Ile-du Cap-Breton			
	Sans erreur d'échantillonnage		Avec erreur d'échantillonnage		Sans erreur d'échantillonnage		Avec erreur d'échantillonnage	
Paramètre	Estimation	T	Estimation	T	Estimation	T	Estimation	T
Alpha (1)	-0.296	-3.23	0.862	2.08	-0.260	-2.85	-0.231	-0.68
Sigma	0.0597	-	0.0032	-	0.1049	-	0.0520	-
Tendance	0.00427	1.01	0.00420	1.89	0.00607	0.79	0.00598	1.50
Janvier	0.064	3.60	0.048	1.93	-0.007	-0.23	-0.003	-0.10
Février	0.083	4.80	0.078	3.30	0.027	0.89	0.028	0.97
Mars	0.166	10.20	0.165	6.40	0.171	5.76	0.164	5.76
Avril	0.106	6.60	0.104	4.10	0.099	3.33	0.089	3.19
Mai	0.009	0.60	0.016	0.70	-0.008	-0.28	-0.007	-0.24
Juin	-0.101	-6.00	-0.088	-3.30	-0.029	-0.96	-0.033	-1.17
Juillet	-0.016	-1.20	-0.014	-0.63	0.082	2.77	0.081	3.13
Août	-0.058	-3.60	-0.062	-2.37	-0.011	-0.37	-0.009	-0.30
Septembre	-0.106	-6.60	-0.105	-3.96	-0.104	-3.51	-0.098	-3.18
Octobre	-0.081	-4.80	-0.071	-3.08	-0.084	-2.83	-0.069	-2.44
Novembre	-0.026	-1.80	-0.029	-1.08	-0.063	-2.10	-0.074	-2.46

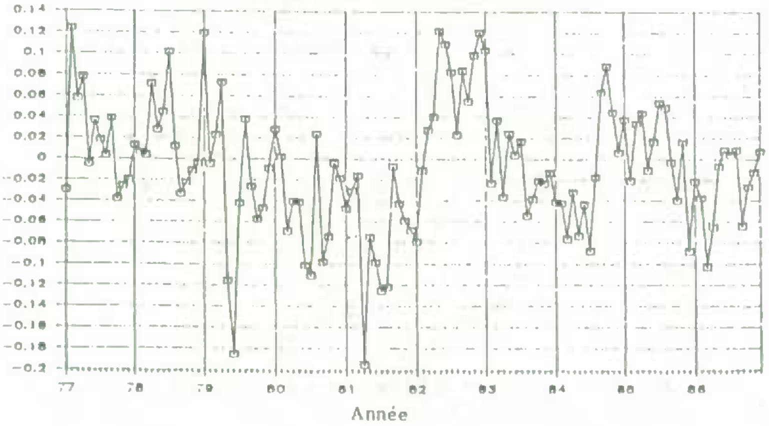
Graphique 1a

Nombre de chômeurs, Nouvelle Écosse, 1977-1986



Graphique 1b

Nombre de chômeurs, Nouvelle Écosse, 1977-1986
Valeurs observées - Lissées



Les résultats de la Nouvelle-Écosse présentent quelques similitudes avec celles du Cap-Breton. Les estimations de régression des modèles "avec erreur" et "sans erreur" sont assez proches. Remarquons que le niveau de signification des estimations par régression du modèle "avec erreur" est beaucoup plus faible que dans celui "sans erreur". La réduction de la variance pour le modèle "avec erreur" par rapport à celui "sans erreur" est beaucoup plus forte que celle entre les deux mêmes modèles pour les données de l'Ile du Cap-Breton. Cependant, le résultat le plus intéressant se trouve dans la composante AR. Les deux modèles montrent que cette dernière est significative pour chaque modèle. Les estimations, par contre, sont très différentes. Le modèle "sans erreur" donne une estimation de α égale à 0.296. Le modèle "avec erreur" donne une estimation de α égale 0.862. Il est évident que les interprétations que l'on pourrait en faire seront différentes. Intuitivement, après élimination de la tendance des effets mensuels, on pourrait s'attendre à ce que l'estimation du mois précédent ait une corrélation positive avec celle du mois courant. C'est exactement le cas du modèle "avec erreur". Il semblerait que la composante AR négative estimée pour le modèle "sans erreur" absorbe une partie du processus des erreurs d'enquête, ce qui se traduit par une interprétation erronnée des données.

Le graphique 1a montre les estimations lissées calculées à partir du modèle incorporant les erreurs d'enquêtes surimposées sur les données originales de la Nouvelle-Écosse. Le graphique 2a montre des estimations lissées semblables pour l'Ile du Cap-Breton. Les valeurs observées moins les estimations lissées des séries de la Nouvelle-Écosse figurent au graphique 1b. On peut voir que la récession de 1981 a une incidence importante. Avant 1981, les estimations lissées avaient tendance à être plus élevées que les valeurs originales, alors qu'après 1981, les estimations lissées ont tendance à être inférieures aux valeurs originales. Les estimateurs observés moins les estimateurs lissés de l'Ile du Cap-Breton figurent au graphique 2b. Ils semblent suivre une allure plus aléatoire que les chiffres pour la Nouvelle-Écosse, probablement à cause des erreurs d'échantillonnage plus importantes des données pour l'Ile du Cap-Breton.

En résumé, lorsque la composante de l'erreur d'échantillonnage est incorporée, le meilleur modèle peut différer de celui dans lequel on laisse de côté l'échantillonnage, ou il peut donner une interprétation complètement différente du modèle. Les données de l'Ile du Cap-Breton présentent une situation pour laquelle l'erreur d'échantillonnage est considérée, où un modèle de régression va expliquer de façon satisfaisante les données, tandis que le modèle qui laisse de côté le processus d'enquête doit inclure une composante AR. Par contre, les données de la Nouvelle-Écosse nécessitaient une composante AR pour les deux modèles, mais donnaient des interprétations complètement différentes pour ses composantes. D'autres travaux seront nécessaires pour évaluer les modèles concurrents. En particulier, comme les erreurs de prédiction d'un pas en avant peuvent être combinées avec les estimations pour donner un processus normal indépendant, il est possible d'évaluer ces prédictions en utilisant les procédures de l'analyse résiduelle habituelle. Des travaux ultérieurs vont présenter en détail les résultats de l'incorporation de cette analyse.

BIBLIOGRAPHIE

Ansley, C.F., and Kohn, R. 1985. A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. *Journal of Statistical Computation and Simulation*. 21: 135-169.

Ansley, C.F., and Kohn, R. 1986. Prediction mean squared error for state space models with estimated parameters. *Biometrika*. 73: 467-473.

Binder, D.A., and Dick, J.P. 1989. Modelling and estimation for repeated surveys. *Survey Methodology*. 14: 29-46.

Blight, B.J.N., and Scott, A.J. 1973. A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B. 35: 61-68.

Harvey, A.C., and Phillips, G.D.A. 1979. Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*. 66: 49-58.

Jones, R.G. 1979. The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*. 21: 45-56.

Jones, R.G. 1980. Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, Series B. 42: 221-226.

Kohn, R., and Ansley, C.F. 1986. Estimation, prediction and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*. 81: 751-761.

Lee, H. 1987. Estimation of panel correlations for the Canadian Labour Force Survey. *Technical Report* SSMD-89-023E. Statistics Canada.

Scott, A.J., and Smith, T.M.F. 1974. Analysis of repeated surveys using time series methods. Journal of the American Statistical Association. 69: 674-678.

Scott, A.J., Smith, T.M.F., and Jones, R.G. 1977. The application of time series methods to the analysis of repeated surveys. *International Statistics Review*. 45: 13-28.

Tunnicliffe-Wilson, G. 1989. On the Use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Series B. 51: 15-27.

$$m(T|T) - \hat{m}(T|T) = [\frac{-\partial\hat{m}(T|T)}{\partial\phi}]' (\hat{\phi} - \phi), \tag{4.3}$$

où, ϕ est le vecteur des paramètres inconnus et $\hat{\phi}$ est son estimation. La variance de (4.2) est donc approximativement

$$\begin{aligned} \mathrm{Var}[z_T - \hat{m}(T|T)] &= V_0(T|T) \\ &+ [\frac{\partial\hat{m}(T|T)}{\partial\phi}]' V_\phi [\frac{\partial\hat{m}(T|T)}{\partial\phi}] , \end{aligned} \tag{4.4}$$

où V_ϕ est la matrice de covariance des paramètres inconnus. On estime l'expression (4.4) en utilisant les valeurs des paramètres estimés. C'est la même méthode que celle de Ansley et Kohn (1986).

5. DONNÉES DE L'ENQUÊTE SUR LA POPULATION ACTIVE

Pour illustrer cette procédure, nous avons pris des données de l'enquête sur la population active du Canada (EPA). L'EPA est une enquête par panel rotatif mensuel. Chaque panel, qui contient le sixième des ménages choisis, reste dans l'échantillon pendant six mois consécutifs. Le plan de sondage est un plan de sondage stratifié à degrés multiples. Les unités primaires d'échantillonnage sont retirées après environ deux ans.

Les données étaient le nombre mensuel estimé de chômeurs de janvier 1977 à décembre 1986 en Nouvelle-Écosse et dans la région infraprovinciale de la Nouvelle-Écosse qui correspond à l'Ile du Cap-Breton. On a choisi cette province parce que les erreurs d'échantillonnage étaient modérées par rapport aux provinces plus grandes, et aussi, parce qu'on disposait de données infraprovinciales. Le logarithme des données de la Nouvelle-Écosse figure au graphique 1a, celui des données de l'Ile du Cap-Breton, au graphique 2a. Les modèles ont été ajustés selon cette série transformée.

Lee (1987) a estimé les autocorrélations des erreurs de l'enquête pour la Nouvelle-Écosse avec un retard de 11. Grâce à ces autocorrélations, nous avons utilisé la méthode des moments pour estimer les coefficients de τ^2, $\phi(B)$ et $\psi(B)$ figurant dans (2.7). On a trouvé un bon ajustement grâce à un modèle ARMM (3,6). Les paramètres estimés étaient $\phi_1 = 0.2575$, $\phi_2 = -0.358$, $\phi_3 = -0.6041$, $\psi_1 = -0.1847$, $\psi_2 = -0.5873$, $\psi_3 = 0.3496$, $\psi_4 = 0.0647$, $\psi_5 = 0.0982$, $\psi_6 = 0.0347$, et $\tau^2 = 0.7246$. Les k_t de (2.6) étaient les erreurs types estimées des estimations, calculés en prenant une approximation obtenue du développement ensérie de Taylor des logarithmes.

On a ajusté une série de modèles aux données en supposant qu'il n'y avait pas d'erreurs d'échantillonnage, c'est-à-dire lorsque tous les k_t étaient égaux à zéro. On a ensuite rajusté ces modèles en utilisant la structure posée par hypothèse pour l'erreur d'enquête. Nous avons comparé les valeurs des paramètres estimés. Tout comme lorsqu'on suppose que la structure d'erreur d'enquête est non nulle, nous avons calculé les valeurs lissées des estimations de l'enquête et nous avons comparé leurs erreurs-types aux erreurs-types des séries originales.

Au début, le modèle sélectionné pour les séries de la Nouvelle-Écosse contenant l'erreur d'enquête était un ARMMI saisonnier $(1,1,0)$ $(0,0,1)_{12}$) avec un terme de régression déterministe pour tenir compte de la saisonnalité. Les 12 variables de régression comprenaient un terme linéaire et une variable auxiliaire pour chacun des 11 premiers mois. La variable auxiliaire d'un mois de référence prenait la valeur 1 pour le mois de référence, - 1 pour décembre et 0 pour les autres mois. Remarquons qu'un terme de coordonnée à l'origine n'est pas estimable parce que les différences premières des données sont ajustées. Les paramètres estimés de ce modèle étaient très instables et on a décidé de laisser tomber la composante de la moyenne mobile saisonnière. Le modèle devenait un modèle ARMMI $(1,1,0)$ avec un terme de régression déterminé. On a utilisé le même modèle pour les données de la Nouvelle-Écosse sans tenir compte de l'erreur de sondage et des données pour le Cap-Breton.

Les estimations des paramètres de la Nouvelle-Écosse et de l'Ile du Cap-Breton figurent au Tableau 1. Nous présentons des estimations qui ne tiennent pas compte de la composante de l'erreur d'enquête dans les colonnes "sans erreur d'échantillonnage". Les estimations des deux modèles pour l'Ile du Cap-Breton, en particulier les estimations par régression, sont très semblables. A noter que la composante AR a également des estimations semblables et que le modèle "avec erreur d'échantillonnage" a réduit sensiblement la variance. La colonne intitulée "valeurs de T" présente la statistique des tests en supposant une valeur vraie de zéro pour le paramètre. A noter que le niveau de signification des estimations par régression est assez proche dans chaque cas. Toutefois, le modèle "sans erreur d'échantillonnage" donne un niveau de signification élevé ($t = -2.85$) pour les composantes AR(1), tandis que le modèle incorporant l'erreur d'enquête n'a pas à inclure la composante AR dans le modèle ($t = -0.68$). Ces résultats ont même été acceptés dans un modèle de règlement pour les séries du chômage dans l'Ile du Cap-Breton pour le modèle avec le processus d'enquête incorporé. Si on laisse de côté l'erreur d'enquête, le modèle inclurait un terme rattachant l'estimation du mois précédant à celle du mois courant.

Comme nous pouvons représenter chacune des composantes de y_t dans l'expression (2.1) par un modèle à espace d'états, il est plus simple de combiner les modèles individuels dans un modèle général en prolongeant le vecteur d'états pour inclure les vecteurs d'états des composantes. L'équation des observations est alors la somme des trois composantes individuelles.

4. ESTIMATIONS DU MODÈLE A ESPÀCE D'ÉTATS

4.1 Estimation des paramètres

Les paramètres inconnus de modèle sont σ^2, et les coefficients de $\lambda(A)$, $\alpha(A)$, $\nu(A)$ et $\beta(A)$. Nous effectuons les itérations sur log (σ^2), plutôt que sur σ^2, afin d'éviter les problèmes des valeurs négatives. A noter que les coefficients de régression γ sont inclus comme paramètres des vecteurs d'états. Le modèle du vecteur des observations $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ de la section 3 équivaut à

$$\mathbf{y} = \mathbf{M}\eta + \zeta, \tag{4.1}$$

où η est j-varié $N(0, \kappa I)$, et ζ T-varié $N(0, \mathbf{W})$ et $\mathbf{M}$ est un matrice $T \times j$.

Kohn et Ansley (1986) recommandent de minimiser la limite de $\kappa^{j/2}$ fois la fonction de vraisemblance des données lorsque κ tend vers l'infini. On peut montrer que la limite de la fonction de vraisemblance équivaut à la fonction de vraisemblance marginale de $\mathbf{y} - \mathbf{M}\hat{\eta}$, où $\hat{\eta}$ est l'estimation du maximum de vraisemblance de η lorsque $\mathbf{M}$ et $\mathbf{W}$ sont connus. Tunnicliffe-Wilson (1989) ont montré que le jacobien de la transformation des données $\mathbf{y}$ en $(\hat{\eta}, \mathbf{y} - \mathbf{M}\hat{\eta})$ ne dépend des paramètres de modèles de $\mathbf{W}$ une fois que $\mathbf{M}$ est connu. Par ailleurs, la dérivée de la transformation de $\mathbf{y}$ en $\hat{\eta}$ est $\mathbf{M}$. Ansley et Kohn (1985) ont montré que M ne dépend pas des paramètres inconnus. En utilisant le filtre de Kalman modifié, les calculs de la fonction de vraisemblance marginale sont simples.

La procédure que nous avons utilisée calcule à la fois la fonction de vraisemblance marginale et ses dérivées premières par rapport aux paramètres inconnus. Il s'agit de prendre les dérivées premières des conditions initiales et de $\mathbf{m}(t|t')$ et des composantes de $\mathbf{V}(t|t')$ pour $t=t'$ et $t=t'+1$. Tous les calculs ont été faits en utilisant PROC IML de SAS.

On a maximisé la fonction de vraisemblance en prenant une modification de la méthode du scoring. Cette modification prévoit l'utilisation de pas différents. A chaque itération, on a calculé la fonction de vraisemblance au pas précédent, ainsi qu'au même pas, multiplié et divisé par une constante déterminée à l'avance. (Nous avons utilisé 1.1 comme facteur). L'étape suivante était de maximiser les fonctions de vraisemblance parmi les trois points. Chaque fois, on a vérifié si les paramètres étaient inclus dans l'intervalle. On y est arrivé en vérifiant si la matrice de covariance initiale du vecteur d'états était semi- définie positive. Si le pas n'était pas inclu dans l'intervalle, on le divisait une nouvelle fois par la constante et on recommençait tous les calculs.

Afin d'obtenir la matrice de variance estimée des paramètres estimés, on a utilisé l'inverse de l'information de Fisher. Ce calcul est facile, puisque les dérivées premières de la fonction de vraisemblance existent.

4.2 Estimation des valeurs lissées

On peut obtenir les valeurs lissées des estimations en posant comme étant égale à zéro la composante du vecteur d'états qui correspond à l'erreur d'enquête. Toutefois, la question de l'estimation de sa variance demeure. Pour calculer l'erreur type de l'estimation lissée, il faut tenir compte du fait que les paramètres inconnus ont été estimés à partir des données, en particulier lorsque la série est courte (voir Jones (1979)).

Pour obtenir la variance de $\mathbf{g}'\mathbf{z}_t$, il suffit de calculer la variance $z_T - \hat{\mathbf{m}}(T|T)$, où $\hat{\mathbf{m}}(T|T)$ est l'estimation de $\mathbf{m}(T|T)$ aux valeurs des paramètres estimées. En effet, le vecteur d'états a été augmenté afin d'inclure $\mathbf{g}'\mathbf{z}_t$. Or,

$$\begin{aligned} z_T - \hat{\mathbf{m}}(T|T) &= [z_T - \mathbf{m}(T|T)] \\ &\quad + [\mathbf{m}(T|T) - \hat{\mathbf{m}}(T|T)]. \end{aligned} \tag{4.2}$$

La première composante du membre droit de (4.2) a comme variance conditionnelle $\mathbf{V}(T|T) = \mathbf{V}_0(T|T)$, en supposant que $\mathbf{V}_1(T|T) = 0$. La deuxième composante de (4.2) est un terme de biais et est indépendante de la première, puisqu'elle ne dépend que des données $\mathbf{y}$. En développant en série de Taylor le deuxième membre autour des valeurs des paramètres réels et en laissant de côté les termes plus élevés, nous avons la deuxième composante de (4.2)

3.2 Modèle de θ

Harvey et Phillips (1979) ont présenté une méthode pour mettre le modèle ARMMI (2.4) sous la forme de l'espace d'états donné par (3.1). La dimension de z_t est $r = \max(p+d+sP+sD,\ q+sQ)$. En augmentant $A + (A_1, \ldots, A_{p+d+sP+sD})$ ou $b = (b_1, \ldots, b_{q+sQ})$ en y ajoutant des zéros pour obtenir la dimension r, on peut récrire le modèle ARMMI sous la forme donnée en (3.1), où $h_t' = (1, 0, \ldots, 0)$, $G_t' = (1, -b_1, \ldots, -b_{r-1})$ et

$$F = \left|\begin{array}{c|c} A_1 & \\ \vdots & I_{r-1} \\ A_{r-1} & \\ \hline A_r & 0' \end{array}\right|,$$

où I_{r-1} est la matrice identité $(r-1) \times (r-1)$ et 0' est un vecteur ligne de zéros.

Dans cette formule, le vecteur d'états $z_t = (z_{1t}, \ldots, z_{rt})'$ est défini par

$$\begin{aligned} z_{it} &= A_i\theta_{t-1} + A_{i+1}\theta_{t-2} + \ldots + A_r\theta_{t-(r-i+1)} \\ &\quad - b_{i-1}\varepsilon_t - b_i\varepsilon_{t-1} - \ldots - b_{r-1}\varepsilon_{t-(r-i)}, \end{aligned} \qquad (3.5)$$

pour $i = 2, 3, \ldots, r$ and $z_{1t} = \theta_t$.

Pour compléter la spécification de $\{\theta_t\}$, il faut les conditions initiales de z_0. Elles figurent dans Ansley et Kohn (1985), dont un résumé est donné ci-dessous.

De par l'expression (2.5), $\{u_t\}$ est un processus ARMM. Nous définissons

$$\theta\_ = (\theta_0, \theta_{-1}, \ldots, \theta_{-S})',$$

où $S = \max(0,\ p+sP+d+sD-1)$. Nous posons

$$u\_ = (u_0, u_{-1}, \ldots, u_{-R})',$$

où $R = \max\ (0,\ p+sP-1)$. Enfin, nous posons

$$w\_ = (\theta_{-R-1}, \theta_{-R-2}, \ldots, \theta_{-S})',$$

où $S > R$.

Or, on suppose que $u\_$ est un processus ARMM stationnaire, de sorte que l'on peut calculer sa matrice de covariance à partir de l'expression (2.5). On suppose que $w\_$ est $N(0, \kappa\ I)$ et indépendant de $u\_$. Comme $(u\_', w\_')'$ est une combinaison linéaire de $\theta\_$, il est possible de calculer la matrice de covariance de θ. Si l'on prend la formule de l'expression (3.5) pour z_0, il est possible de calculer la matrice de covariance initiale. Remarquons que lorsque d et D sont égaux à zéro, de sorte qu'il n'y a pas de différenciation dans le modèle, alors $w\_$ est le vecteur nul, et nous avons $u\_ = \theta\_$.

3.3 Modèle pour les données observées

A la section 2 nous avons supposé que $e_t = k_t\,\omega_t$, où ω_t est un modèle ARMM (m,n). Il ressort donc de la discussion dans la section 3.3. que e_t peut être représenté sous une forme d'espace d'états avec $h_t = (k_t, 0, \ldots, 0)'$ et $e_t = h'_t z_t$.

On peut de même représenter la composante de régression. Supposons $z_0 = \gamma$, qui sont les coefficients de régression, avec par hypothèse une moyenne nulle et une covariance κI. L'équation de transition est simplement $z_{t+1} = z_t$.

$\psi(B)$ peuvent être estimés directement à partir des données de l'enquête en utilisant des méthodes de plan. Cependant, en général, les autres paramètres sont inconnus. Ces paramètres comprennent γ, σ^2, et les coefficients de $\lambda(A)$, $\alpha(A)$, $\nu(A)$ et $\beta(A)$. Les x_t du terme de régression sont supposés connus.

3. FORMULATION PAR ESPACE D'ÉTATS DU MODÈLE

3.1 Formulation générale

Le modèle décrit à la Section 2 peut être formulé comme un modèle d'espace d'états avec une distribution a priori partielle. Ceci offre un certain nombre d'avantages. Par l'utilisation d'un filtre de Kalman modifié, le modèle permet le calcul d'une fonction de vraisemblance marginale qui peut être maximisée pour estimer les paramètres inconnus. Il peut également lisser les estimations de l'enquête originale en éliminant les estimations des erreurs d'enquête des données.

Dans le modèle à espace d'états, deux processus se produisent simultanément. Le premier, le système des observations, donne de façon détaillée la façon dont les observations dépendent de l'état actuel des paramètres du processus. Le deuxième, le système de transition, montre en fait comment les paramètres évoluent dans le temps.

Pour les modèles à espace d'états que nous considérons ici, l'équation des observations s'écrit sous la forme

$$y_t = \mathbf{h}_t' z_t \qquad (3.1a)$$

et l'équation de transition est

$$z_t = F z_{t-1} + G\xi_t, \qquad (3.1b)$$

où z_t est un vecteur d'états $(r\times 1)$ et $\mathbf{h}_t$ est un vecteur fixe $(r\times 1)$. Dans l'équation de transition, F est une matrice de transition fixe $(r\times r)$, $\mathbf{G}$ est une matrice fixe $(r\times m)$ et les ξ_t sont des vecteurs normaux indépendants de moyenne zéro et de covariance $\mathbf{U}$.

La condition finale pour compléter la spécification du processus à espace d'états est la condition initiale de z_0. Ici, nous allons utiliser la formulation des conditions impropres de Kohn et Ansley (1986). En général, nous supposons que z_0 est une distribution normale à r variables partiellement diffuse, de moyenne $\mathbf{m}(0|0) = 0$ et de matrice de covariance $\mathbf{V}(0|0)$, où

$$\mathbf{V}(0|0) = \kappa \mathbf{V}_1(0|0) + \mathbf{V}_0(0|0) \qquad (3.2)$$

pour κ grand.

Nous dénotons la moyenne conditionnelle de z_t compte tenu des observations jusqu'à la période t' inclusivement par $\mathbf{m}(t|t')$, et la variance conditionnelle par $\mathbf{V}(t|t')$, où

$$\mathbf{V}(t|t') = \kappa \mathbf{V}_1(t|t') + \mathbf{V}_0(t|t'). \qquad (3.3)$$

Les formules de récurrence pour $t=t'$ et pour $t=t'+1$ se trouvent dans Kohn et Ansley (1986). Ces auteurs l'appelent le filtre de Kalman modifié.

Comme le modèle de $\{y_t\}$ donné par (2.1) contient des erreurs d'enquête $\{e_t\}$, une estimation des composantes sans erreur d'enquête, donnée par la formule

$$y_t \text{ (smoothed)} = x_t'\gamma + \theta_t \qquad (3.4)$$

est souvent intéressante. Lorsque le membre droit de (3.4) peut être exprimé comme $\mathbf{g}'_t z_t$, pour un $\mathbf{g}_t'$, il est alors possible d'obtenir la moyenne et la variance conditionnelle de la combinaison linéaire $\mathbf{g}'_t z_t$ pour toutes les données grâce au filtre de Kalman modifié. Pour cela, les calculs de récurrence sont appliqués jusqu'au temps t pour obtenir $\mathbf{m}(t|t)$ et $\mathbf{V}(t|t)$. Le vecteur d'états z_t est augmenté de l'état $z_{t,\ r+1} = \mathbf{g}'_t z_t$, et $\mathbf{m}(t|t)$ et $\mathbf{V}(t|t)$ sont également augmentés de façon appropriée. La matrice F dans (3.1b) est modifiée par l'ajout de l'équation $z_{t+1,\ r+1} = z_{t,r+1}$. Après ces modifications, le filtre de Kalman modifié peut être utilisé comme auparavant, de sorte que la dernière composante de $\mathbf{m}(t|t)$ donne l'espérance conditionnelle de $\mathbf{g}'_t z_t$, pour toutes les données $y_1, y_2, \ldots y_T$. De plus, la dernière composante diagonale de $\mathbf{V}(t|t)$ donne la variance conditionnelle. Il est possible de généraliser cette procédure pour y inclure un nombre quelconque d'estimations lissées et leurs covariances conditionnelles.

Nous décrivons d'abord un modèle à moyennes mobiles autorégressives saisonnières intégrées de $\{\theta_t\}$. Soit B l'opérateur arrière; $\nabla = 1-B$ et $\nabla_s = 1-B^S$, ou s est la période saisonnière. Nous définissons les fonctions polynomiales suivantes:

$$\lambda(A) = 1 - \lambda_1 A - \lambda_2 A^2 - \dots - \lambda_P A^P,$$

$$\alpha(A) = 1 - \alpha_1 A - \alpha_2 A^2 - \dots - \alpha_p A^p,$$

$$\nu(A) = 1 - \nu_1 A - \nu_2 A^2 - \dots - \nu_Q A^Q,$$

et

$$\beta(A) = 1 - \beta_1 A - \beta_2 A^2 - \dots - \beta_q A^q.$$

Le modèle ARMMI saisonnier $(p,d,q)(P,D,Q)_s$ pour $\{\theta_t\}$ est donné par

$$\lambda(B^S)\alpha(B)\nabla^d\nabla_s^D\theta_t = \nu(B^S)\beta(B)\varepsilon_t, \tag{2.2}$$

où les ε_t sont indépendants $N(0,\sigma^2)$. Nous définissons $a(B) = \lambda(B^S)\alpha(B)$ $(p+sP)$, un polynôme de degré un polynôme de degré $(q+SQ)$, $\Delta(B) = \nabla^d\nabla_s^D$ un polynôme de degré $(d+sD)$; $b(B) = \nu(B^S)\beta(B)$ un polynôme de degré $(q+SQ)$, $A(B) = a(B)\Delta(B)$ un polynôme de degré $(p+d+sP+sD)$-degree polynomial; $u_t = \Delta(B)\theta_t$, un processus ARMM $(p+sP,q+sQ)$. D'autres expressions de (2.2) sont par conséquent:

$$a(B)\Delta(B)\theta_t = b(B)\varepsilon_t, \tag{2.3}$$

$$A(B)\theta_t = b(B)\varepsilon_t, \tag{2.4}$$

et

$$a(B)u_t = b(B)\varepsilon_t, \tag{2.5}$$

Examinons maintenant les erreurs d'enquête $\{e_t\}$ de l'expression (2.1). On supposera que les tailles des échantillons de l'enquête répétée sont suffisamment importantes pour que les erreurs des estimations d'enquête puissent être approximées par une distribution normale multivariée. Dans le cas le plus simple où les enquêtes ne se chevauchent pas et où les fractions de sondage sont petites, on peut supposer que les e_t sont indépendantes. Dans une enquête par panel rotatif, les erreurs d'enquête sont habituellement corrélées. Dans ce cas, comme les corrélations entre les occasions d'enquête sont nulles après la rotation des panels, on peut utiliser un processus de moyennes mobiles pur pour décrire le processus des erreurs d'enquête.

Sinon, si on remplace l'échantillon aléatoire d'unités à chaque enquête, un processus autorégressif pur pourrait le mieux décrire le processus. Des modèles plus compliqués sont également possibles. Ainsi, dans un plan de sondage à deux degrés, une partie des unités du premier degré peuvent être remplacées de façon aléatoire à chaque occasion, tandis que les unités de deuxième degré pourraient avoir un plan de sondage par panel rotatif. On peut représenter ceci par un processus de moyennes mobiles autorégressives.

Nous supposons ici que le processus des erreurs d'enquête est donné par la formule

$$e_t = k_t\omega_t, \tag{2.6}$$

où $\{\omega_t\}$ est un processus ARMM (m,n) défini par

$$\phi(B)\omega_t = \psi(B)\eta_t \tag{2.7}$$

et

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_m B^m,$$

et

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_n B^n,$$

Les η_t sont indépendants, $N(0,\tau^2)$. On a inclus le facteur k_t dans (2.6) pour tenir compte des variances non homogènes, même lorsque la fonction d'autocorrélation est homogène dans le temps.

Dans le module que nous venons juste de décrire, nous supposons que τ^2, les k_t et les coefficients de $\phi(B)$ et de

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ANALYSE DES MODÈLES ARMMI SAISONNIERS AU MOYEN DES DONNÉES D'ENQUÊTE

D.A. Binder et J.P. Dick[1]

RÉSUMÉ

Le modèle saisonnier ARMMI est souvent utilisé dans l'analyse des modèles de séries chronologiques. Cependant, les erreurs d'enquête des données originales sont souvent laissées de côté dans l'analyse. Nous montrons, grâce à des modèles à espace d'états avec des conditions initiales en partie non appropriées, comment on estime les paramètres inconnus de ce modèle en employant les méthodes du maximum de vraisemblance. Par ailleurs, il est possible de lisser les estimations d'enquête en utilisant pour cela une méthode bayésienne empirique. Nous utilisons ces techniques dans le cas d'une série de données relatives au chômage provenant de l'Enquête sur la Population Active.

1. INTRODUCTION

Il est de pratique courante d'utiliser des séries chronologiques afin d'analyser des données semblables recueillies à plusieurs occasions dans le temps. La plupart des méthodes habituelles de ces analyses supposent que les données ont été observées sans erreur, ou qu'elles ont des erreurs de mesures indépendantes. Toutefois, dans l'analyse des données d'enquêtes répétées, lorsqu'il y a chevauchement des unités d'échantillonnage entre les différentes occasions, les erreurs d'enquête peuvent ainsi être corrélées dans le temps.

Un modèle habituellement utilisé dans l'analyse des séries chronologiques est le modèle de régression saisonnier à moyenne mobile autorégressive intégrée (ARMMI), que nous examinerons ici. Nous montrons comment incorporer des erreurs d'enquête (peut-être corrélées) dans l'analyse. En particulier, nous examinons le cas de l'erreur (de plan) d'enquête que l'on peut considérer comme un processus ARMMI à une constante multiplicative près.

Lorsque l'on retient un tel modèle du comportement des caractéristiques de la population, il est possible de calculer l'erreur quadratique moyenne minimum, ou, ce qui est équivalent, l'estimateur linéaire bayésien de la caractéristique à un moment donné. Cet estimateur incorpore la structure du modèle que les estimateurs classiques, tels que les estimateurs linéaires sans biais à variance minimum, laissent de côté. Lorsque les paramètres du modèle sont estimés à partir des données simples d'enquête, les estimateurs sont des estimateurs bayésiens empiriques.

Blight et Scott (1973), Scott et Smith (1974), Scott, Smith et Jones (1977), Jones (1980) et d'autres ont examiné les conséquences d'utiliser certains modèles stochastiques pour les moyennes de la population dans le temps. Dans Binder et Dick (1989), ces résultats ont été généralisés par l'emploi de modèles à espace d'états et de filtres de Kalman. Dans cette communication, nous prolongeons le cadre de façon à inclure le modèle où la différenciation de la série originale des moyennes de la population donne un modèle ARMM. Nous utilisons la méthode du filtre de Kalman modifiée de Kohn et Ansley (1986). Pour estimer les paramètres inconnus, nous maximisons la fonction du maximum de vraisemblance par la méthode du scoring. Cette méthode permet également de manipuler de façon simple les données manquantes. Nous montrons également comment des estimations de l'enquête peuvent ainsi être lissées pour incorporer les caractéristiques du modèle grâce aux méthodes bayésiennes empiriques. Nous donnons également les intervalles de confiance de ces valeurs lissées, en utilisant pour cela la méthode décrite par Ansley et Kohn (1986). Un exemple de ce modèle est présenté à la section 5, utilisant les données de chômage provenant de l'enquête sur la population active du Canada. Cet exemple montre les conséquences sur les estimations des paramètres du modèle lorsque l'on tient compte des erreurs d'enquête. Nous calculons également une estimation lissée du processus sous- jacent en vertu des hypothèses du modèle.

2. LE MODÈLE

Supposons que nous ayons une série d'estimations ponctuelles provenant d'une enquête répetée d'une caractéristique de la population, donnée par y_1, y_2, ..., y_T. Nous supposons que y_t peut être décomposé en trois composantes, de sorte que

$$y_t = x'_t\gamma + \theta_t + e_t, \tag{2.1}$$

où $x'_t\gamma$ est un terme de régression déterminé, θ_t est un paramètre de la population suivant un modèle de séries chronologiques, et e_t est l'erreur d'enquête, dont l'espérance est par hypothèse zéro.

[1] D.A. Binder, Division des méthodes d'enquêtes - entreprises, J.P. Dick, Division des méthodes d'enquêtes sociales, Statistique Canada, Parc Tunney, Ottawa, (Ontario), Canada K1A 0T6

Griliches, Z. (1971), "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change", dans *Price Index and Quality Change*, Ed. Zvi Griliches, Harvard University Press, pp. 55-87.

Harvey, A.C. (1981), *"Time Series Models,"* Philip Allan, Deddington, Oxford.
(1984, "A Unified View of Statistical Forecasting Procedures," *Journal of Forecasting*, 3, 245-275.

Harvey, A.C. et Phillips, G.D.A. (1979), "Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances," *Biometrika*, 66, 49-58.

Hofsten, E.V. (1952), *"Price Indexes and Quality Change,"* Bokforlaget Forum AB, Stockholm.

Hsiao, C. (1974), "Statistical Inference for a Model with Both Random Cross- Sectional and Time Effects," *International Economic Review*, 15, 12-30.

Johnson, L.W. (1977), "Stochastic Parameter Regressions: An Annotated Bibliography," *International Statistical Review*, 45, 257-272.

-- (1980), "Stochastic Parameter Regression: An Additional Annotated Bibliography," *International Statistical Review*, 48, 95-102.

LaMotte, L.R., et McWhorter, A. (1977), "Estimation, Testing and Forecasting with Random Coefficient Regression Models," dans *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, pp. 814-817.

Maddala, G.S. (1977), *"Econometrics,"* McGraw-Hill, Kogakusta.

Meinhold, R.J. et Singpurwalla, N.D. (1983), "Understanding the Kalman Filter," *The American Statistician*, 37, 123-127.

Pfeffermann, D., et Nathan, G. (1981), "Regression Analysis of Data from a Cluster Sample," *Journal of the American Statistical Association*, 76, 681-689.

Pfeffermann, D., et Smith, T.M.F. (1985), "Regression Models for Grouped Populations in Cross-Section Surveys," *International Statistical Review*, 53, 37-59.

Rosenberg, B. (1973a), "The Analysis of Cross-Section of Time Series by Stochastically Convergent Parameter Regression," *Annals of Economic and Social Measurement*, 2, 399, 428.

-- (1973b) "A Survey of Stochastic Parameter Regression," *Annals of Economic and Social Measurement*, 2, 381-397.

Schweppe, F. (1965), "Evaluation of Likelihood Functions for Gaussian Signals," *IEEE Transactions on Information Theory*, 11, 61-70.

Stone, R. (1956), *"Quality and Price Indexes in National Accounts,"* Organisation européenne de coopération économique, Paris.

Swamy, P.A.V.B., et Mehta, J.S. (1977), "Estimation of Linear Models with Time and Cross-Sectionally Varying Coefficients," *Journal of the American Statistical Association*, 72, 890-898.

Watson, M.W., et Engle, R.F. (1983), "Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models," *Journal of Econometrics*, 23, 385-400.

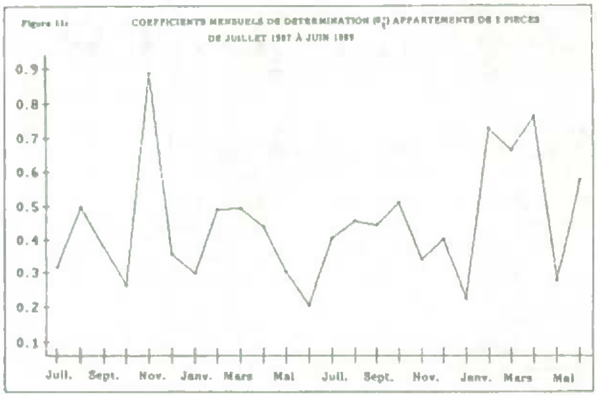


Figure 11: COEFFICIENTS MENSUELS DE DETERMINATION (R_t^2) APPARTEMENTS DE 3 PIECES DE JUILLET 1987 À JUIN 1989

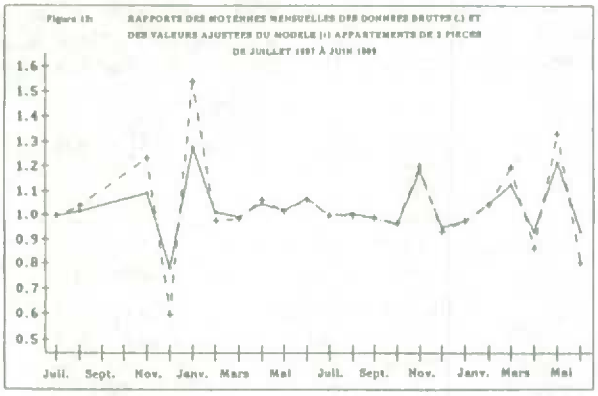


Figure 12: RAPPORTS DES MOYENNES MENSUELLES DES DONNEES BRUTES (.) ET DES VALEURS AJUSTEES DU MODELE (+) APPARTEMENTS DE 3 PIECES DE JUILLET 1987 À JUIN 1989

Comme le montre la figure 11, les statistiques R^2 sont dans la plupart des cas supérieures à 0.4, ce qui est très élevé pour ce genre de données. A titre de comparaison, les valeurs obtenues lorsque les coefficients ont été estimés à l'aide des MCO n'étaient que de 10 à 20% supérieures. (Les MCO des R^2 sont le maximum possible pour un ensemble donné de variables de régression.) La figure 12 fait ressortir une étroite correspondance entre les rapports mensuels des données brutes et les rapports des valeurs ajustées. Il importe de souligner que ces rapports ne sont pas des estimations de l'IPL puisqu'ils ne sont pas nécessairement fondés sur les prix de maisons de qualité similaire. Toutefois, il est très encourageant de constater que les rapports des valeurs après ajustement sont à ce point proches des rapports des données originales.

7. CONCLUSION

Les résultats de cette étude montrent qu'il est possible d'estimer de façon efficace les rapports de régression à l'intérieur de petites cases en utilisant un modèle tenant comte de la variation des coefficients de régression dans le temps. Evidemment, des tests plus poussés devront être effectués en vue de confirmer la validité du modèle. Nous sommes déjà en train de mettre à l'essai l'ensemble du modèle défini par les équations (3.1)-(3.3) et prenant aussi en compte les modifications relatives à la robustesse proposées dans la section 5. Lorsque nous comparerons les résultats de la présente étude avec ceux de l'essai du modèle complet avec et sans les modifications, nous serons en mesure de mieux juger de l'efficacité du modèle et de ces modifications. Nous prévoyons aussi tester la qualité de l'ajustement du modèles pour ce qui est de prédire les prix de vente des maisons enregistrés après publication de l'indice. Le fait que les dates de comptabilisation des ventes n'ont pas été codées dans nos fichiers de travail actuels explique pourquoi nous n'avons pas encore réalisé un tel test.

REMERCIEMENTS

Cette étude a été réalisée au moment où le premier auteur cité bénéficiait d'un congé sabbatique de l'Hebrew University. Les auteurs tiennent à remercier Statistique Canada pour sa contribution à l'étude dans le cadre de son programme de bourses en recherche.

BIBLIOGRAPHIE

Adelman, I., et Griliches, Z. (1961), "On an Index of Quality Change," *Journal of the American Statistical Association*, 56, 535-548.

Anderson, B.O.D., et Moore, J.B. (1979), "*Optimal Filtering*", Prentice-Hall, Englewood Cliffs, N.J.

Ansley, C.F., et Kohn, R. (1986), "Prediction Mean Squared Error for State Space Models with Estimated Parameters," *Biometrika*, 73, 467-473.

Battese, G.E., Harter, R.M., et Fuller, W.A. (1988), "An Error-Componenets Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Castles, I. (1987), "The Australian Consumer Price Index Treatment of Home Ownership Costs," document d'information n° 6441.0 au catalogue, Australian Bureau of Statistics, Belconen ACT 2616.

Cooley, T.F., et Prescott, E.C. (1976), "Estimation in the Presence of Stochastic Parameter Variation," *Econometrica*, 44, 167-184.

Court, A.T. (1939), "Hedonic Price Indexes with Automotive Examples," dans *The Dynamics of Automobile Demand*, pp. 99-117. New York: General Motors Corporation.

Dielman, T.E. (1983), "Pooled Cross-Sectional and Time Series Data: A Survey of Current Statistical Methodology," *The American Statistician*, 37, 111- 122.

L'instabilité des estimateurs des MCO, comparativement à celle des estimateurs lissés et filtrés, ressort encore plus du tableau 1 qui suit où sont comparées les variances des MCO et des estimateurs lissés pour les mois d'avril à mai 1989. Comme on pouvait s'y attendre, les estimateurs lissés qui sont fondés sur les données de l'ensemble des mois de référence ont des variances beaucoup plus petites dans tous les cas.

Tableau 1: Variances des MCO et des estimateurs lissés des coefficients de régression

Mois	Estimateur	Ordonnées	Superficie	Âge	District 1	District 2
Avril 89	MCO	.174	.064	.006	.011	.011
	Lissés	.068	.025	.0002	.002	.0017
Mai 89	MCO	.471	.142	.0021	.011	.010
	Lissés	.093	.033	.0003	.002	.0013

La faible variation d'un mois à l'autre des estimateurs filtrés et lissés peut laisser sous-entendre que les équations de régression sont pratiquement fixes dans le temps. Nous avons déjà souligné que les variances des résidus des coefficients de régression se sont révélées hautement significatives, signe qu'un modèle qui permet aux coefficients de régression de varier avec le temps est plus approprié. Afin de mieux démontrer ce point, nous comparons, dans les figures 9 et 10, les moyennes et les EQM des erreurs de prédictions obtenues à la suite de l'utilisation des estimateurs filtrés (voir figures 4 et 6) et des estimateurs globaux des MCO fondés sur l'ensemble des données jusqu'au temps t inclusivement. Les lignes tracées dans ces figures sont très révélatrices puisqu'on peut voir que des coefficients de régression fixes, c'est-à-dire qui ne varient pas dans le temps, entraînent des biais de prédiction importants et croissants qui à leur tour augmentent les EQM de prédiction.

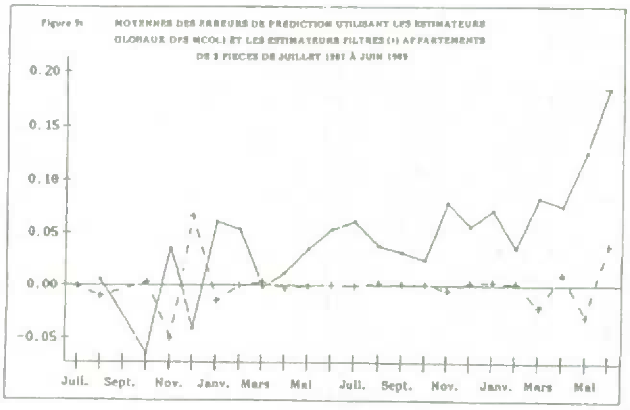


Figure 9: MOYENNES DES ERREURS DE PRÉDICTION UTILISANT LES ESTIMATEURS GLOBAUX DES MCO(-) ET LES ESTIMATEURS FILTRÉS (+) APPARTEMENTS DE 3 PIÈCES DE JUILLET 1987 À JUIN 1989

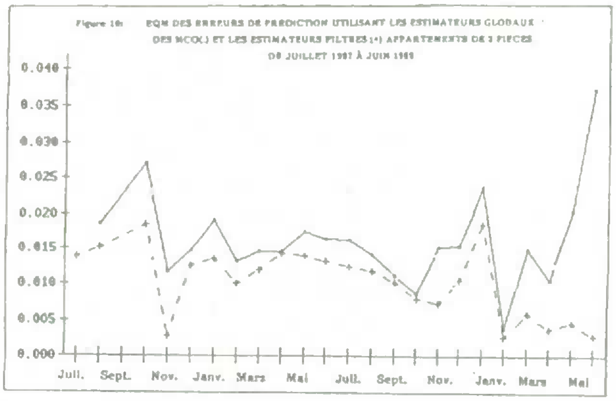
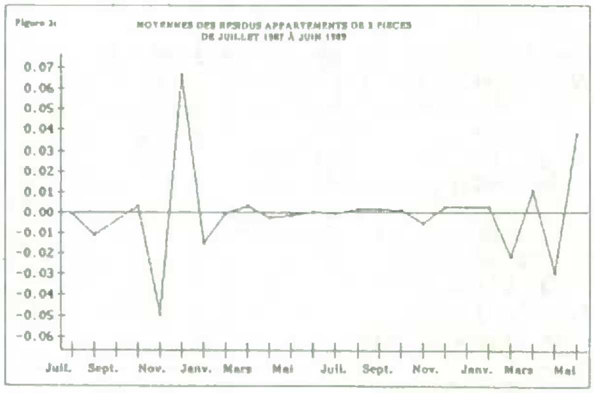


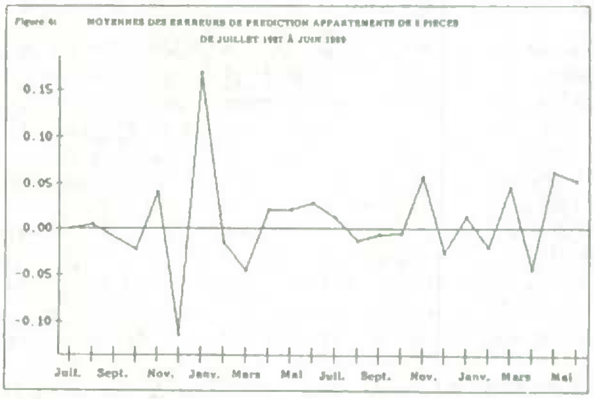
Figure 10: EQM DES ERREURS DE PRÉDICTION UTILISANT LES ESTIMATEURS GLOBAUX DES MCO(-) ET LES ESTIMATEURS FILTRÉS (+) APPARTEMENTS DE 3 PIÈCES DE JUILLET 1987 À JUIN 1989

La question la plus importante concernant la qualité de l'ajustement du modèle est sa capacité d'estimer les IPL. Pour répondre en partie à la question, nous avons calculé deux ensembles de statistiques: i) les "coefficients de détermination" mensuels (R^2) définis par

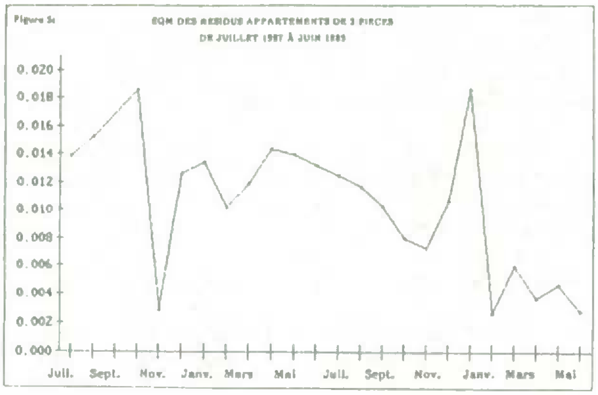
$$R_t^2 = 1 - \{ \sum_{j=1}^{n_t} [y_{tj} - \exp(\underset{\sim}{x}'_{tj} \underset{\sim}{\alpha}_t^F)]^2 / \sum_{j=1}^{n_t} (Y_{tj} - \bar{Y}_t)^2 \}$$

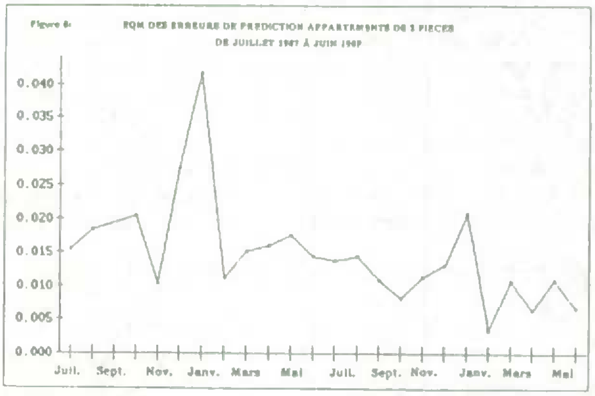
où $\bar{Y}_t$ est la moyenne des prix de vente au cours du mois t (résultats tracés dans la figure 11); et ii) les rapports des moyennes mensuelles des données brutes, $R^r_{t|t-1} = \bar{Y}_t / \bar{Y}_{t-1}$, et des moyennes des valeurs ajustées correspondantes $R^f_{t|t-1} = \bar{f}_t / \bar{f}_{t-1}$ ou $\bar{f}_t = \sum_{j=1}^{n_t} \exp(\underset{\sim}{x}'_{tj} \hat{\underset{\sim}{\alpha}}{}_t^F) / n_t$. Les deux séries de rapports sont illustrées graphiquement dans la figure 12. Il est à noter que toutes les statistiques ci-dessus ont été calculées après avoir été ramenées de l'échelle logarithmique à l'échelle normale.



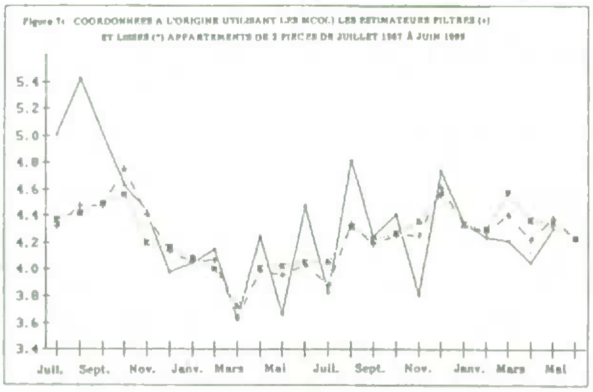


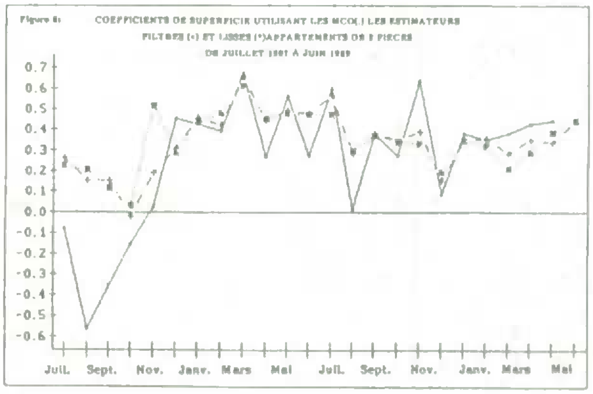
Les figures 3 et 4 tracent les moyennes mensuelles des résidus et des erreurs de prédiction pour la période de juillet 1987 à juin 1989. Les figures 5 et 6 tracent les EQM qui y correspondent. Soulignons que les données des 12 derniers mois n'ont pas été utilisées pour l'estimation des variances du modèle. Comme on pouvait s'y attendre, les erreurs de prédiction sont plus variables que les résidus, mais il n'y a rien dans les quatre figures qui fassent ressortir des défaillances systématiques du modèle et les résultats obtenus pour les mois de juillet 1987 à juin 1988 (les données relatives à ces mois ont été utilisées dans le processus d'estimation) sont semblables aux résultats pour les douze autres mois. Il est à noter qu'étant donné que presque tous les coefficients des VMQ suivent un modèle de marche aléatoire, $\underset{\sim}{\hat{\alpha}}_{t|t-1} = \underset{\sim}{\hat{\alpha}}^{F}_{t-1}$ de sorte que, par exemple, la moyenne résiduelle négative relativement importante observée pour le mois de novembre 1987 s'accompagne d'une importante erreur de prédiction négative pour le mois de décembre 1987.





Les figures 7 et 8 montrent les estimateurs mensuels des ordonnées à l'origine et des coefficients de superficie obtenus au moyen des moindres carrés ordinaires (MCO) utilisant les données des mois correspondants uniquement, les estimateurs filtrés et les estimateurs lissés. Comme on peut le constater, les estimateurs filtrés et lissés sont dans l'ensemble très semblables (ils sont évidemment beaucoup plus distancés dans les premiers mois dont les données ne sont pas tracées sur le graphique) et ils ne varient que très peu d'un mois à l'autre. Par contre, l'estimateur des MCO affiche une variation importante d'un mois à un autre et pour les mois de juillet 1987 et d'octobre 1987, les estimateurs des coefficients de superficie sont même négatifs.





6. RÉSULTATS EMPIRIQUES

Afin de démontrer que le modèle pouvait être appliqué aux prix d'achat des maisons en Israël, nous avons procédé à un ajustement distinct du modèle pour chacune des cinq cases de la ville de Jérusalem en nous servant à cette fin des données d'observation se rapportant aux ventes effectuées entre les mois de septembre 1982 et de juin 1988. Les cases sont définies en fonction du nombre de pièces - lequel va de 1 à 5. Par manque de temps et pour d'autres raisons techniques, nous n'avons pas encore introduit simultanément des données de cases différentes, de sorte que le modèle se fonde uniquement sur les rapports chronologiques entre les coefficients de régression des cases, tels que définis par l'équation (3.2). Comme nous avons utilisé les données relatives à une seule case à chacune des étapes du traitement, nous n'avons pas non plus introduit les modifications analysées dans la section 5. Nous faisons présentement l'essai d'un programme informatique détaillé qui permet d'ajuster l'ensemble du modèle défini par les équations (3.1) - (3.3), qui est fondé sur les méthodes d'estimation décrites dans la section 4 et qui introduit les modifications de la section 5. Les résultats pourront être communiqués aux lecteurs intéressés, sur demande, tout comme les données brutes.

L'ajustement du modèle aux cinq cases a donné des résultats assez uniformes du point de vue de l'importance des estimateurs de variance du modèle. Aussi, sauf dans le cas des appartements de 5 pièces, la variance $\hat{\delta}_s^2$ du coefficient de la pente était non significative, ce qui suppose un modèle de marche aléatoire pour le coefficient des ordonnées à l'origine étant donné que le coefficient initial de la pente avait été fixé à zéro. Pour les appartements de 5 pièces, $\hat{\delta}_s^2 = 4 \times 10^{-4}$, ce qui est très faible quoique significatif au niveau 0.5%. De même, sauf dans le cas des appartements de 4 pièces, les variances des ordonnées à l'origine et des quatre autres VMQ définies par l'équation (2.1) se sont aussi révélées très significatives, ce qui confirme notre hypothèse de départ voulant que les coefficients de régression varient de façon stochastique avec le temps. Pour ce qui est des appartements de 4 pièces, la variance du coefficient des ordonnées à l'origine est une fois de plus très significative, la variance du coefficient de la pente aréolaire est significative au niveau 10%, mais les estimateurs de variance restants ne sont pas significatifs.

(Dans la section 3, nous laissions entendre qu'en raison des taux d'inflation mensuels relativement élevés et constants en Israël, les coefficients des ordonnées à l'origine pourraient croître de façon linéaire. Cependant, nous supposions également que les autres coefficients seraient constants dans le temps, ce qui n'est évidemment pas le cas. Il semble aussi que l'IPL est beaucoup plus variable que l'indice général des prix à la consommation.)

Dans le reste de cette section, nous présentons plusieurs graphiques qui illustrent le rendement du modèle lorsqu'appliqué à des appartements de 2 pièces. Nous nous sommes limités aux appartements de 2 pièces tout simplement pour des raisons d'espace et de toutes façons, les résultats relatifs aux autres cases sont généralement très semblables. Nous utilisons les notations et définitions suivantes:

Y_{tj}^{*} - le logarithme du prix de vente de l'appartement j au cours du mois t, $j=1\ldots n_t$, $t=1,2\ldots T$

$\underset{\sim}{x}_{tj}$ - les VMQ correspondant à l'appartement j au cours du mois t. Les VMQ sont les ordonnées à l'origine et les quatre variables déterminées par l'équation (2.1) (à l'exclusion de la fonction de temps $g_2(t)$)

$\hat{\underset{\sim}{\alpha}}_t^{OLS}$ - les estimateurs des MCO des coefficients des VMQ fondés sur les ventes réalisées au cours du mois t.

$\hat{\underset{\sim}{\alpha}}_t^{F}$ - les estimateurs filtrés des coefficients des VMQ fondés sur les ventes réalisées jusqu'au mois t inclusivement (équation 4.3)

$\hat{\underset{\sim}{\alpha}}_t^{S}$ - les estimateurs lissés fondés sur l'ensemble des ventes réalisées au cours de tous les mois (équation 4.4)

$\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{F}$ - les valeurs prévues des coefficients des VMQ une période à l'avance

$e_{tj} = (Y_{tj}^{*} - \underset{\sim}{x}_{tj}' \hat{\underset{\sim}{\alpha}}_t^{F})$ - les résidus observés concernant la vente j du mois t

$m_t = \sum_{j=1}^{n_t} e_{tj}/n_t$ et $mse_t = \sum_{j=1}^{n_t} e_{tj}^2/n_t$ - les moyennes mensuelles et les EQM des résidus

$e_{ptj} = (Y_{tj}^{*} - \underset{\sim}{x}_{tj}' \hat{\underset{\sim}{\alpha}}_{t|t-1})$ - l'erreur de prédiction associée à la vente (tj)

$m_{pt} = \sum_{j=1}^{n_t} e_{ptj}/n_t$ et $mse_{pt} = \sum_{j=1}^{n_t} e_{ptj}^2/n_t$ - les moyennes mensuelles et les EQM des erreurs de prédiction.

5.2 Estimation robuste à l'aide des équations augmentées

Dans la section 5.1, nous proposions de modifier les équations du modèle (3.1) en imposant la série de liaisons (5.1), ce qui a pour effet de garantir la robustesse des estimateurs de régression et de les protéger contre des changements soudains de valeur des coefficients. Pour faciliter les calculs, nous augmentons les vecteurs $\underset{\sim}{Y}_t$ de l'équation (4.1) par les facteurs d'échelle $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ $[i=1,2\cdots I(t)$ désignant le nombre de liaisons au temps $t]$, nous augmentons les matrices Z_t par les vecteurs des rangées correspondantes $(w_{t1}^{(i)} 1'_{nt1} Z_{t1} \cdots w_{tK}^{(i)} 1'_{ntK} Z_{tK})$ et nous fixons à zéro les variances respectives des termes résiduels. La série augmentée d'équations avec l'équation (4.2) forment un pseudo-modèle d'espace d'états qui peut être estimé à l'aide des équations du filtre de Kalman (4.3). Soulignons que la pseudo-matrice des V-C $\Sigma_t^{(P)}$ du vecteur résiduel augmenté n'est plus définie positive (les dernières rangées et colonnes de $I(t)$ sont constituées de zéros), mais cela ne pose pas de problèmes de calcul.

L'inconvénient de l'application du filtre de Kalman au pseudo-modèle est que les matrices des V-C des estimateurs de régression ne réussissent pas à tenir compte de la variabilité réelle des moyennes globales des données brutes. S'il est vrai que, comme nous venons de le voir dans la section 5.1, cette variabilité peut ne pas être prise en compte lorsque les moyennes sont fondées sur un nombre suffisamment important de ventes, une procédure à la fois meilleure et plus robuste consiste à modifier la formule de mise à jour de la matrice des V-C P_t (équation 4.3) de façon à ce que les variances et covariances des variables aléatoires $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ soient prises en compte. Supposons que $\underset{\sim}{Y}_t^{(A)}$ et $Z_t^{(A)}$ représentent le vecteur augmenté Y et la matrice Z au moment t et désignons par $\Sigma_t^{(A)}$ la matrice des V-C réelles des termes résiduels $[\underset{\sim}{Y}^{(A)} - Z_t^{(A)} \underset{\sim}{\alpha}_t]$. La matrice $\Sigma_t^{(A)}$ est de l'ordre $[n_t + I(t)]$ où Σ_t occupe les n_t premières lignes et colonnes parmi lesquelles on trouve également les variances et covariances des moyennes $\sum_k w_{tk}^{(i)} \sum_j Y_{tkj}$ et où le vecteur $\underset{\sim}{Y}_t$ se situe dans les lignes et colonnes restantes. Représentant par $\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}$ le prédicteur robuste de $\underset{\sim}{\alpha}_{t-1}$ obtenu au temps $(t-1)$ à l'aide du pseudo-modèle et par $P_{t-1}^{(A)}$ la matrice des V-C réelles des erreurs $(\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_{t-1})$, l'estimateur des états modifiés au moment t est obtenu de la manière suivante

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + P_{t|t-1}^{(A)} \; Z_t^{(A)'} (F_t^{(P)})^{-1} \; [\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} T \hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}] \tag{5.2}$$

où $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)} \; T' + \Lambda)$ et $F_t^{(P)} = [Z_t^{(A)} P_{t|t-1}^{(A)} \; Z_t^{(A)'} + \Sigma_t^{(P)}]$. (À comparer avec (4.3)). On peut voir que la matrice des V-C réelles $P_t^{(A)}$ des erreurs $(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t)$ satisfait l'équation récursive

$$P_t^{(A)} = [I - K_t^{(P)} Z_t^{(A)}] \; P_{t|t-1}^{(A)} + K_t^{(P)} [\Sigma_t^{(A)} - \Sigma_t^{(P)}] K_t^{(P)'} \tag{5.3}$$

où $K_t^{(P)} = P_{t|t-1}^{(A)} \; Z_t^{(A)'} \; (F_t^{(P)})^{-1}$ est le pseudo-gain de Kalman. La première expression du côté droit de l'équation (5.3) correspond à la formule habituelle de mise à jour du filtre de Kalman (à comparer avec (4.3)). La seconde expression est un facteur de correction qui tient compte des variances et covariances réelles des moyennes $\Sigma_k \; w_{tk}^{(i)} \sum_j Y_{tkj}$, non prises en compte dans la première expression.

Le filtre de Kalman modifié défini par les équations (5.2) et (5.3) produit les prédicteurs robustes $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ à la place des prédicteurs dépendants du modèle optimal, mais utilise les matrices des V-C exactes du modèle. Par conséquent, ce filtre peut servir à l'estimation de routine des vecteurs des coefficients et lorsque le modèle se vérifie, les résultats seront similaires à ceux obtenus avec le filtre optimal. Pour les périodes où le modèle ne se tient pas, la formule de mise à jour (5.3) peut être inexacte (selon le genre de défaillance du modèle), mais les prédicteurs $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ satisferont quand même les liaisons linéaires (5.1). Les équations de lissage (4.4) et la matrice des V-C de (4.5) peuvent être modifiées en fonction d'une utilisation semblable des prédicteurs robustes.

où $\underset{\sim}{\lambda}_{(i-1)}$ est l'estimateur de $\underset{\sim}{\lambda}$, tel qu'obtenu dans la (i-1)ième itération, $I[\underset{\sim}{\lambda}_{(i-1)}]$ est la matrice d'information évaluée à $\underset{\sim}{\lambda}_{i-1}$ et $g[\underset{\sim}{\lambda}_{(i-1)}]$ est le gradient de la probabilité logarithmique évaluée à $\underset{\sim}{\lambda}_{i-1}$. Le coefficient r_i est la longueur de pas variable introduite pour garantir que $L[\underset{\sim}{\lambda}_{(i)}] \geq L[\underset{\sim}{\lambda}_{(i-1)}]$ dans chaque itération. La valeur de r_i est déterminée par une recherche par quadrillage. Les formules du k-ième élément du vecteur de gradient et du $k^{ième}$ élément de la matrice d'information sont données dans Watson et Engle (1983).

Une fois les variances et covariances du modèle estimées, elles peuvent être remplacées par les paramètres vrais des équations du filtre de Kalman (4.3) - (4.5) pour produire les estimateurs des coefficients de régression et des matrices des V-C dont nous avons besoin. Il est à noter que les matrices des V-C ne tiennent pas compte de la variabilité supplémentaire attribuable à la nécessité d'estimer les éléments contenus dans $\underset{\sim}{\lambda}$. Ansley et Kohn (1986) proposent des facteurs de correction de l'ordre 1/t* pour tenir compte de cette variation supplémentaire dans un modèle d'espace d'états.

On a conçu un programme informatique permettant d'appliquer les méthodes décrites dans cette section pour l'estimation du filtre de Kalman en se servant à cette fin de la procédure PROC-IML du SAS.

5. MODIFICATIONS SERVANT DE PROTECTION CONTRE LES DÉFAILLANCES DU MODÈLE

5.1 Description du problème et modifications proposées

Le recours à un modèle pour le calcul de l'IPL est inévitable en raison du problème des changements de qualité. Il faut cependant aussi se protéger contre les défaillances du modèle. Comme il n'est pas réaliste de vérifier le modèle chaque fois que de nouvelles données deviennent disponibles, il importe à la place de concevoir un "mécanisme intégré" permettant de garantir la robustesse des indices en cas de défaillance du modèle.

Le problème se pose principalement pour les mois où les prix font un bond imprévu. En Israël par exemple, les taux de dévaluation de la monnaie peuvent parfois atteindre jusqu'à 10%. Bien que la dévaluation s'accompagne le plus souvent de mesures strictes de contrôle des prix visant à geler les anciens prix, ces politiques ont peu d'effet sur les prix d'achat des maisons qui sont déterminés par les négociations directes entre acheteurs et vendeurs et qui, par conséquent, échappent aux contrôles. Par ailleurs, le modèle proposé dans la section (3) se fonde sur les rapports qualité/prix passés pour renforcer l'estimation des rapports actuels et, de cette façon, ne s'ajuste qu'avec un certain décalage à des changements aussi soudains.

Afin de remédier au problème, nous proposons de modifier les estimateurs de régression calculés aux diverses périodes de façon à ce qu'ils obéissent à certaines liaisons linéaires obtenues par égalisation des moyennes globales des données brutes et des valeurs prévues dans le modèle. Plus précisément, nous proposons de compléter les équations du modèle (3.1) par des liaisons linéaires de la forme

$$\sum_k W_{tk}^{(i)} \left(n_{tk}\gamma_{tk} + 1_{ntk}{}'X_{tk}\underset{\sim}{\beta}_{tk}\right) = \sum_k W_{tk}^{(i)} \sum_j Y_{tkj} \qquad \begin{array}{l} i=1,2\cdots I(t) \\ t=1\cdots T \end{array} \tag{5.1}$$

où les coefficients $\{W_{tk}^{(i)}\}$ sont des poids fixes corrigés pour satisfaire $\sum_k n_{tk} W_{tk}^{(i)} = 1$. Il importe de souligner que les liaisons (5.1) ne représentent pas l'information externe au sujet des valeurs possibles des coefficients de régression. Elles jouent plutôt le rôle de système de contrôle pour garantir que les estimateurs du modèle s'ajustent plus rapidement aux changements soudains de comportement des coefficients de régression. Ainsi, les variances des estimateurs de régression modifiés sont légèrement supérieures aux variances des estimateurs optimaux du modèle. Evidemment, lorsqu'il ne se produit aucun changement du genre et que les variances des moyennes globales sont suffisamment petites, on peut s'attendre que les liaisons soient approximativement satisfaites sans avoir à les imposer de façon explicite. L'idéal serait d'introduire plusieurs liaisons distinctes pour chaque période à l'étude, mais il est essentiel que les variances des moyennes globales correspondantes soient suffisamment petites pour assurer que des modifications soient quand même nécessaires et qu'elles ne nuisent pas à la fluctuation aléatoire des données brutes.

Au nombre des moyennes globales qui peuvent être utilisées dans le cas des données sur l'achat des maisons, on compte: i) moyennes distinctes de toutes les données comprises dans les cases comptant un nombre élevé de ventes; ii) moyennes distinctes des données de combinaisons de cases comptant un nombre donné de pièces; et iii) moyennes distinctes des données de cases comptant un nombre différent de pièces, par exemple de toutes les données relatives à une ville donnée. A noter qu'en raison des corrélations entre les coefficients de régression des diverses cases, une liaison appliquée à un sous-ensemble de cases modifiera les estimations de régression de toutes les cases. Battese, Harter et Fuller (1988) proposent sensiblement le même genre de modification dans le contexte de l'estimation de données régionales.

$$\hat{\underset{\sim}{\alpha}}_t = \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1} (\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$$

$$P_t = (I - P_{t|t-1} Z_t' F_t^{-1} Z_t) P_{t|t-1} \qquad (4.3)$$

où $\hat{\underset{\sim}{Y}}_{t|t-1} = Z_t \hat{\underset{\sim}{\alpha}}_{t|t-1}$ est le MELSB de $\underset{\sim}{Y}_t$ au moment (t-1) de sorte que $\underset{\sim}{e}_t = (\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$ est le vecteur d'innovations avec comme matrice des V-C $F_t = (Z_t P_{t|t-1} Z_t' + \Sigma_t)$.

Les nouvelles données observées au moment t peuvent aussi servir pour la mise à jour (le lissage) des estimateurs passés. Désignant par t* le mois le plus récent pour lequel on dispose d'observations, le lissage se fait à l'aide des équations

$$\hat{\underset{\sim}{\alpha}}_{t|t\star} = \hat{\underset{\sim}{\alpha}}_t + P_t T' P^{-1}_{t+1|t}(\hat{\underset{\sim}{\alpha}}_{t+1|t\star} - T\hat{\underset{\sim}{\alpha}}_t)$$

$$P_{t|t\star} = P_t + P_t T' P^{-1}_{t+1|t}(P_{t+1|t\star} - P_{t+1|t}) P^{-1}_{t+1|t} T P_t \;;\; t=2,\ 3,\ \ldots\ t\star \qquad (4.4)$$

où $P_{t|t\star}$ est la matrice des V-C des erreurs de prédiction $(\hat{\underset{\sim}{\alpha}}_{t|t\star} - \underset{\sim}{\alpha}_t)$. À noter que $\hat{\underset{\sim}{\alpha}}_{t\star|t\star} = \hat{\underset{\sim}{\alpha}}_{t\star}$ et que $P_{t\star|t\star} = P_{t\star}$ ce qui définit les valeurs de départ nécessaires aux équations de lissage.

Lorsque nous utilisons le modèle pour estimer l'IPL d'un mois donné t, nous devons estimer les vecteurs $\underset{\sim}{\alpha}_t$ et $\underset{\sim}{\alpha}_{t-1}$. Afin d'estimer la variance de l'indice estimé, nous devons estimer la matrice des covariances entre les estimateurs $\hat{\underset{\sim}{\alpha}}_t$ and $\hat{\underset{\sim}{\alpha}}_{t-1|t}$. La matrice des covariances a la forme

$$E(\hat{\underset{\sim}{\alpha}}_t - \underset{\sim}{\alpha}_t)(\hat{\underset{\sim}{\alpha}}_{t-1|t} - \underset{\sim}{\alpha}_{t-1})' = (I - P_{t|t-1} Z_t' F_t^{-1} Z_t) T P_{t-1} \qquad (4.5)$$

4.3 Estimation des matrices des V-C et initialisation du filtre

L'application actuelle du filtre de Kalman exige l'estimation des éléments inconnus des matrices Σ_t et Λ ainsi que l'initialisation du filtre, c'est-à-dire l'estimation du vecteur $\underset{\sim}{\alpha}_o$ et de la matrice correspondante des V-C P_o des erreurs d'estimation. Dans cette section, nous décrivons brièvement les méthodes d'estimation utilisées dans la présente étude.

Les paramètres inconnus du modèle ont été estimés à partir de la théorie du maximum de vraisemblance. En supposant une distribution normale pour les termes résiduels $\underset{\sim}{\varepsilon}_t$ et $\underset{\sim}{\eta}_t$ et une loi à priori diffuse pour $\underset{\sim}{\alpha}_o$, la fonction logarithmique de probabilité des observations $\underset{\sim}{Y}_3 \ldots \underset{\sim}{Y}_t$ dépend de $\underset{\sim}{Y}_1$ et de $\underset{\sim}{Y}_2$ et peut s'écrire

$$L(\underset{\sim}{\lambda}) = \text{constante} - \frac{1}{2} \sum_{t=3}^{T} (\log |F_t| + \underset{\sim}{e}_t' F_t^{-1} \underset{\sim}{e}_t) \qquad (4.6)$$

où $\underset{\sim}{\lambda}$ contient les variance et covariances inconnues du modèle écrites sous forme vectorielle. L'équation (4.6) est obtenue par décomposition des erreurs de prédiction, voir Schweppe (1965) et Harvey (1981) pour plus de détails. Pour des matrices données Σ_t et Λ, les innovations $\underset{\sim}{e}_t$ et les matrices des C-V F_t sont obtenues à la suite de l'application des équations du filtre de Kalman (4.3).

Pour pouvoir calculer la fonction de probabilité, il faut initialiser le filtre de Kalman; c'est ce que nous avons fait en utilisant la méthode proposées par Harvey et Phillips (1979). Avec cette méthode, la loi à priori diffuse $\underset{\sim}{\alpha}_o$ se réalise à la suite de l'initialisation du filtre au moment t=0 avec $\underset{\sim}{\alpha}_o = \underset{\sim}{0}$ et $P_o = N \times I$ où N est un nombre fini très grand et I la matrice unité de l'ordre approprié.

Pour maximiser la fonction de probabilité (4.6), nous avons appliqué la méthode de notation avec une longueur de pas variable. Si $\underset{\sim}{\lambda}_{(o)}$ définit les estimations initiales des éléments inconnus contenus dans $\underset{\sim}{\lambda}$, la méthode de notation consiste à résoudre par itération la série d'équations

$$\underset{\sim}{\lambda}_{(i)} = \underset{\sim}{\lambda}_{(i-1)} + r_i \{I[\underset{\sim}{\lambda}_{(i-1)}]\}^{-1} g[\underset{\sim}{\lambda}_{(i-1)}] \qquad (4.7)$$

$\underset{\sim}{\alpha}'_{tK} = (\gamma_{tK}, s_{tK}, \underset{\sim}{\beta}'_{tk})$ les coefficients de régression correspondant à la case k et supposons que $\underset{\sim}{\alpha}'_t = (\underset{\sim}{\alpha}'_{t1} \ldots \underset{\sim}{\alpha}'_{tK})$.

En nous fondant sur la notation ci-dessus, la série des équations définies par (3.1) peut être exprimée sous la forme abrégée suivante

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\varepsilon}_t \; ; \quad E(\underset{\sim}{\varepsilon}_t) = \underset{\sim}{0} \quad , E(\underset{\sim}{\varepsilon}_t \underset{\sim}{\varepsilon}_t') = \Sigma_t \qquad (4.1)$$

où $\Sigma_t = \text{Diag}\ [\sigma_1^2 1'_{nt1} \ldots \sigma_K^2 1'_{ntK}]$.

Soit $T^* = \begin{bmatrix} 1,1 & 0 \\ 0,1 & \\ 0 & I_{m-2} \end{bmatrix}$ une matrice quasi-diagonale de l'ordre mxm où I_{m-2} est la matrice unité de l'ordre (m-2) et soit $T = I_K \otimes T^*$ où $\otimes$ représente le produit de kronecker.

Le système d'équations définies par (3.2) et (3.3) peut être exprimée sous la forme abrégée suivante

$$\underset{\sim}{\alpha}_t = T\underset{\sim}{\alpha}_{t-1} + \underset{\sim}{n}_t \quad ; \quad E(\underset{\sim}{n}_t) = \underset{\sim}{0} \quad , \quad E(\underset{\sim}{n}_t \underset{\sim}{n}'_t) = \Lambda \qquad (4.2)$$

où $\underset{\sim}{n}'_t = (\underset{\sim}{n}'_{t1} \ldots \underset{\sim}{n}'_{tK})$ et $\Lambda = [\Lambda_{k\ell},\ k,\ell = 1 \ldots K]$ et

$$\Lambda_{kk} = E(\underset{\sim}{n}_{tk}\ \underset{\sim}{n}'_{tk}) = \begin{bmatrix} \delta_Y^2 & 0 & {}_\beta\delta'_Y \\ 0 & \delta_S^2 & \underset{\sim}{0}' \\ {}_\beta\delta_Y & \underset{\sim}{0} & \Delta_\beta \end{bmatrix} \quad \text{et} \quad \Lambda_{k\ell} = E(\underset{\sim}{n}_{tk}\ \underset{\sim}{n}'_{t\ell}) = \Delta\emptyset,\ k \neq \ell\ .$$

(Les matrices Δ_{kk} et $\Delta_{k\ell}$ sont de l'ordre mXm).

Les équations (4.1) et (4.2) sont conformes à la formulation classique du modèle d'espace d'états (Harvey, 1984), (4.1) représentant l'équation des observations et (4.2) l'équation du système. L'avantage de la restructuration du modèle sous forme d'espace d'états est que les vecteurs $\underset{\sim}{\alpha}_t$ peuvent alors être estimés plus facilement grâce à l'utilisation du filtre de Kalman. Nous décrivons les principales étapes de cette méthode dans la section suivante.

4.2 Estimation du modèle au moyen du filtre de Kalman

Dans cette section, nous supposons que les matrices V-C Σ_t et Λ sont connues. L'estimation des éléments inconnus de ces matrices est le sujet de la section 4.3. Le filtre de Kalman consiste en une série d'équations récursives qui déterminent comment mettre à jour les estimateurs actuels et passés des vecteurs des états du système (les coefficients de régression $\underset{\sim}{\alpha}_t$ du modèle dans le cas qui nous intéresse) et comment prédire les vecteurs futurs chaque fois que de nouvelles données deviennent disponibles. De plus, le filtre fournit les matrices des V-C des divers estimateurs et prédicteurs. Comme la théorie de filtrage de Kalman est élaborée dans de nombreuses publications (voir par ex. Anderson et Moore, 1979 et Meinhold et Singpurwalla, 1983), nous ne présenterons ici que les équations de base.

Supposons que $\hat{\underset{\sim}{\alpha}}_{t-1}$ soit le meilleur estimateur linéaire sans biais (MELSB) de $\underset{\sim}{\alpha}_{t-1}$ pour toutes les données observées jusqu'au moment (t-1). Etant donné que $\hat{\underset{\sim}{\alpha}}_{t-1}$ est le MELSB de $\underset{\sim}{\alpha}_{t-1}$, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}_{t-1}$ est le MELSB de $\underset{\sim}{\alpha}_t$ au moment (t-1). De plus, si $P_{t-1} = E(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})'$, est la matrice des V-C des erreurs de prédiction au moment (t-1), $P_{t|t-1} = TP_{t-1}T' + \Lambda$ est la matrice des V-C des erreurs de prédiction $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$. (Découle directement de 4.2)

Lorsqu'un nouveau vecteur d'observations devient disponible, le prédicteur de $\underset{\sim}{\alpha}_t$ et la matrice des V-C P_{t-1} sont mis à jour conformément à la formule particulièrement utile lorsque le nombre de cases est restreint est de supposer des corrélations constantes entre résidus applicables à des cases différentes. Ainsi, pour $\underset{\sim}{\eta}'_{tk} = (\eta_{Ytk}, \eta_{stk}, \underset{\sim}{\eta}_{\beta tk})$, cette hypothèse a la forme

$$E(\underset{\sim}{\eta}_{tk}\ \underset{\sim}{\eta}'_{t\ell}) = \Delta\emptyset\,,\ k\neq\ell \tag{3.3}$$

où Δ est une matrice diagonale dont la diagonale principale contient δ^2_Y, δ^2_s ainsi que les éléments diagonaux de Δ_β et où $\emptyset$ est une autre matrice diagonale dont tous les éléments se situent à l'intérieur de l'intervalle (- 1,1). Les éléments diagonaux de $\emptyset$ définissent les corrélations entre les résidus des diverses cases.

Une autre solution possible, lorsqu'une "distance" peut être mesurée entre les diverses cases (par exemple lorsque les cases sont définies en fonction du nombre de pièces comme dans la présente étude), consiste à supposer que les corrélations entre les résidus diminuent à mesure que la distance entre les cases augmente. Cette hypothèse peut s'écrire

$$E(\underset{\sim}{\eta}_{tk}\ \underset{\sim}{\eta}'_{t\ell}) = \Delta\emptyset f(k,\ell);\ k\neq\ell \tag{3.4}$$

où $f(k,\ell)$ est une fonction monotone décroissante des distances $D(k,\ell)$. L'équation (3.3) est un cas particulier évident de (3.4).

ANALYSE: De nombreuses études statistiques et économétriques ont recours à des coefficients de régression stochastiques pour tenir compte des variations longitudinales et (ou) transversales. Johnson (1977,1980) fournit une bibliographie annotée de plus de 150 articles qui traitent de modèles de ce genre. Notre modèle va plus loin que les modèles antérieurs en supposant des tendances linéaires locales pour les ordonnées à l'origine et en structurant les corrélations transversales. Cooley et Prescott (1976) et LaMotte et McWhorter (1977) supposent que tous les coefficients de régression de leur modèle suivent une marche aléatoire, Rosenberg (1973a) suppose des relations autorégressives alors que Hsiao (1974) et Swamy et Mehta (1977) supposent que les coefficients obtenus peuvent être factorisés en deux composantes, soit une moyenne commune et une erreur indépendante, qui permettent de tenir compte des variations longitudinales et transversales. Ces études et de nombreuses autres études sur la régression à l'aide de coefficients stochastiques sont passées en revue et analysées dans Rosenberg (1973b), Maddala (1977, chapitre 7), Dielman (1983) et Pfeffermann et Smith (1985).

Nous avons déjà expliqué, à la fin de la section 2, pourquoi il faut permettre aux coefficients de régression de varier dans le temps. Le modèle de marche aléatoire suppose que les coefficients s'écartent progressivement de leur valeur initiale sans avoir tendance en soi à revenir à une valeur moyenne. Nous estimons que ce genre de modèle convient très bien à l'ajustement des prix d'achat des maisons. Il a également pour avantage de ne comporter qu'un nombre minime de paramètres inconnus, ce qui est très important compte tenu du nombre déjà élevé de paramètres compris dans les équations (3.1) - (3.3).

Le choix de ce modèle particulier pour les ordonnées à l'origine s'est fait en raison des taux d'inflation mensuels relativement élevés en Israël, lesquels ont fluctué autour de 1.5% au cours des deux dernières années. En effet, nous prévoyions que le logarithme des prix de maisons données (la variable dépendante de notre modèle) croîtrait de façon presque linéaire avec le temps, ce qui suppose, pour les valeurs fixes des autres coefficients de régression, que les ordonnées à l'origine définies par les deux premières équations de (3.2) suivraient aussi une tendance sensiblement linéaire.

Le modèle défini par (3.1) - (3.3) comble les lacunes de la méthode actuellement utilisée par le CBS dont il a été question à la fin de la section 2. Les poids attribués aux diverses VMQ ne sont plus constants dans le temps et la fonction de temps déterministe (2.2) est remplacée par une fonction de tendance plus souple qui évolue dans le temps. Les estimateurs calculés pour toute case donnée sont renforcés grâce à l'emprunt de données provenant à la fois de cases voisines et d'observations antérieures. La quantité de telles données est déterminée par la proximité des vecteurs des coefficients (sur une base tant transversale que longitudinale), elle-même décelée par les estimateurs des variances et covariances du modèle (voir section 4 pour plus de détails).

4. ESTIMATION DU MODÈLE

4.1 Représentation du modèle sous forme d'espace d'états

Dans les lignes qui suivent, nous utilisons la notation suivante: nous représentons par $\underset{\sim}{Y}'_t = (\underset{\sim}{Y}'_{t1}\cdots\underset{\sim}{Y}'_{tK})$ le vecteur des observations au moment t de durée $n_t = \sum_{k=1}^{K} n_{tk}$ et par $\underset{\sim}{\varepsilon}'_t = (\underset{\sim}{\varepsilon}'_{t1}\cdots\underset{\sim}{\varepsilon}'_{tK})$ le vecteur correspondant des résidus. Nous supposons que $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, X_{tk}]$ où $\underset{\sim}{0}_{ntk}$ est le vecteur nul de durée n_{tk} et supposons que Z_t est la matrice quasi- diagonale dont le k-ième bloc est compris dans Z_{tk}. La matrice Z_t est de l'ordre $n_t\times(K\times m)$ où m désigne le nombre de colonnes dans chacune des matrices Z_{tk}. Représentons par multiplicative (2.1), on suppose que le rapport entre les prix prévus de maisons de qualité fixe différente demeure constant tout au long des six mois de référence. Comme le marché de l'immobilier est un marché instable qui dépend des négociations entre vendeurs et acheteurs et de la conjoncture économique, il semble plus approprié de laisser les coefficients des VMQ varier dans le temps. (De nombreuses études traitent de la question de l'instabilité des relations économétriques, voir notamment l'analyse faite par Cooley et Prescott, 1976.) Le choix limité de la fonction de temps, bien que fondé sur des données empiriques pour une année en particulier, n'offre pas la souplesse voulue pour tenir compte des changements de prix des maisons d'un mois à un autre. De plus, l'hypothèse associée à la fonction de temps n'est pas assez générale pour qu'elle se vérifie simultanément pour l'ensemble des périodes à l'étude et pour tous les différents types d'habitation. Une autre limite de la méthode est l'interpolation des indices mensuels qui se fait de manière plutôt improvisée.

Le CBS utilise présentement cette méthode pour une raison bien simple: le manque de données au moment du calcul de l'IPL, même dans le cas des cases plus grandes. S'il est vrai qu'on tente d'y remédier en empruntant des données des cases voisines, cela ne résout pas les autres problèmes susmentionnés. Il semble qu'une des principales lacunes de la méthode actuelle est qu'on n'exploite pas les propriétés des séries chronologiques. En effet, les données antérieures à la période de six mois à l'étude ne sont pas prises en compte dans le calcul des indices malgré le fait que ces données ont trait aux mêmes cases et mesurent le même phénomène. Le modèle présenté dans la section suivante tient compte des relations tant longitudinales que transversales qui existent entre les coefficients de régression. En empruntant du passé les renseignements nécessaires, l'estimation des indices peut se faire sur une base mensuelle sans qu'il soit nécessaire de se restreindre à des coefficients fixes pour les VMQ ou à une fonction de temps déterministe, comme c'est le cas avec la méthode actuelle.

3. RÉGRESSION AVEC DES COEFFICIENTS QUI VARIENT DE FAÇON TRANSVERSALE ET LONGITUDINALE

Dans les équations qui suivent, nous représentons par $\underset{\sim}{Y}_{tk}$ le vecteur (n_{tk} x 1) d'observations sur la variable dépendante (les logarithmes des prix de vente dans le cas qui nous intéresse) relatives à la case (domaine) k au moment t, k=1...k, t=1, 2, ... Nous supposons que $\underset{\sim}{Y}_{tk}$ n'est pas vide, bien que l'absence d'observations dans certaines des cases et à certains moments ne pose pas de problème du point de vue méthodologique, comme nous le verrons dans la section 4. Nous supposons en outre que X_{tk} représente la matrice du modèle correspondant (plan) des variables explicatives (les VMQ dans le cas qui nous intéresse). Le modèle de régression dans la case k est défini par

$$\underset{\sim}{Y}_{tk} = \underset{\sim}{1}_{ntk}\, \gamma_{tk} + X_{tk}\, \underset{\sim}{\beta}_{tk} + \underset{\sim}{\varepsilon}_{tk} \; ; \qquad E(\underset{\sim}{\varepsilon}_{tk}) = \underset{\sim}{0},\; E(\underset{\sim}{\varepsilon}_{tk}\, \underset{\sim}{\varepsilon}'_{tk}) = \sigma^2_k\, I_{ntk} \tag{3.1}$$

où $\underset{\sim}{1}_{ntk}$ et I_{ntk} représentent respectivement le vecteur unitaire et la matrice d'unité d'ordre n_{tk}. La principale caractéristique de l'équation (3.1) est que les coefficients γ_{tk} et $\underset{\sim}{\beta}_{tk}$ peuvent varier de façon transversale et longitudinale. Les équations qui suivent précisent la variation dans le temps des coefficients, soit

$$\begin{aligned}
\gamma_{tk} &= \gamma_{t-1,k} + s_{t-1,k} + n_{\gamma tk} \; ; \qquad E(n_{\gamma tk}) = 0,\; E(n_{\gamma tk})^2 = \delta^2_\gamma \\
s_{tk} &= s_{t-1,k} + n_{stk} \; ; \qquad E(n_{stk}) = 0,\; E(n^2_{stk}) = \delta^2_s \\
\underset{\sim}{\beta}_{tk} &= \underset{\sim}{\beta}_{t-1,k} + \underset{\sim}{n}_{\beta tk};\; E(\underset{\sim}{n}_{\beta tk}) = \underset{\sim}{0},\; E(\underset{\sim}{n}_{\beta tk}\, \underset{\sim}{n}'_{\beta tk}) = \Delta_\beta,\; E(\underset{\sim}{n}_{\beta tk}\, n_{\gamma tk}) = {}_{\beta}\underset{\sim}{\delta}_{\gamma}
\end{aligned} \tag{3.2}$$

Nous supposons également que n_{stk} n'est pas corrélé avec ($n_{\gamma tk}$, $\underset{\sim}{n}_{\beta tk}$) et que toutes les corrélations avec décalage sont égales à zéro.

Les équations (3.2) définissent une approximation locale d'une tendance linéaire pour l'ordonnée à l'origine et un modèle de marche aléatoire pour les autres coefficients. Comme les variables explicatives sont habituellement corrélées, les changements de valeur des divers coefficients peuvent aussi être corrélés; on a recours à cette fin à une matrice de variances-covariances (V-C) générale Δ_β (pouvant notamment comprendre des variances résiduelles différentes selon les coefficients) et un vecteur de covariances général ${}_{\beta}\underset{\sim}{\delta}_{\gamma}$.

Une façon simple de prendre en compte les relations transversales entre les coefficients de régression consiste à permettre que les corrélations ne soient pas égales à zéro entre les résidus correspondants des équations (3.2). Toutefois, même avec un petit nombre de cases, il faut donner une certaine structure à ces corrélations si on veut que le nombre de paramètres inconnus du modèle reste raisonnable. Une solution qui semble

$$g_k(t)=\lambda_{k1}t_1 + \lambda_{k2}t_2 \quad \text{où} \quad t_1 = \begin{cases} t & \text{si } t < 4 \\ 3.5 & \text{autrement} \end{cases}, \; t_2 = \begin{cases} 0 & \text{si } t < 4 \\ t-3.5 & \text{autrement} \end{cases} \tag{2.2}$$

Le modèle défini par (2.1) et (2.2) est estimé à l'aide des moindres carrés ordinaires (MCO) donnant les estimations préliminaires $(\hat{\lambda}_{k1}, \hat{\lambda}_{K2})$ avec les variances estimées $\{\hat{V}\ (\hat{\lambda}_{k1}),\ \hat{V}\ (\hat{\lambda}_{k2})\}$.

Etape 2 A l'étape 2, les estimations $(\hat{\lambda}_{k1}, \hat{\lambda}_{K2})$ sont "rétrécies" vers une moyenne commune obtenue des estimateurs calculés pour des cases voisines. Les cases voisines utilisées pour le processus de rétrécissement sont toutes les cases se rapportant à une même ville si les données sont disponibles en nombre suffisant ou, dans le cas contraire, les cases se rapportant à un groupe de villes. Le rétrécissement se fait en considérant les λ-coefficients qui s'appliquent à un groupe donné de cases comme des variables aléatoires indépendantes échangeables, de sorte que

$$E(\lambda_{ka}) = \lambda_a; \quad E(\lambda_{ka} - \lambda_a)\ (\lambda_{\ell b} - \lambda_b) = \begin{cases} \delta_a^2 & a = b,\ k = \ell \\ 0 & \text{autrement} \end{cases} \quad a,b,=1,2 \tag{2.3}$$

Les estimations modifiées et rétrécies sont des estimations empiriques de moindres carrés élargis (Pfeffermann et Nathan, 1981) définies par

$$\hat{\lambda}_{Ka}(e) = G_K\ \hat{\lambda}_{Ka} + (1 - G_K)\ \hat{\lambda}_a(e) \quad ; \quad \hat{\lambda}_a(e) = \sum_K G_K\ \hat{\lambda}_{Ka} / \sum_K G_K \tag{2.4}$$

où $G_K = \hat{\delta}_a^2 / \{\hat{\delta}_a^2 + \hat{V}\ (\hat{\lambda}_{Ka})\}$. Les variances δ_a^2 sont estimées grâce à la procédure itérative proposée par Pfeffermann et Nathan (1981) qui est appliquée à toutes les estimations par case pour tous les groupes, de façon à n'utiliser qu'une seule estimation de variance par trimestre pour chacun des deux coefficients λ. Les λ-coefficients des cases pour lesquelles les données ne sont pas assez nombreuses pour permettre le calcul des estimateurs des MCO sont estimés à l'aide des moyennes correspondantes $\hat{\lambda}_a(e)$, $a=1,2$. Pour faciliter la notation, nous utilisons ci-dessous les symboles $\hat{\lambda}_{Ka}(e)$ pour toutes les cases, que les données soient disponibles ou non.

Etape 3 A l'aide du modèle défini par (2.1) et (2.2), un IPL est estimé pour chacune des cases pour un intervalle de trois mois. L'indice représente la hausse de prix moyenne entre le mois 2 (le point milieu du premier trimestre) et le mois 5 (le point milieu du deuxième trimestre) et il est calculé comme

$$\hat{R}_{K,5/2} = \hat{Y}_{5K.}/\hat{Y}_{2K.} = \exp\ \{1.5\ \hat{\lambda}_{K1}(e) + 1.5\ \hat{\lambda}_{K2}(e)\}$$

où $\hat{Y}_{tK.}$ est le prix prévu (après ajustement) au moment t pour des valeurs moyennes données des VMQ. Soulignons qu'étant donné que nous avons utilisé une relation multiplicative et que nous avons supposé des coefficients fixes pour la période de six mois, le rapport $\hat{R}_{K,5/2}$ est indépendant du choix des valeurs moyennes des VMQ. Il est également à noter qu'en vertu de l'hypothèse de la normalité des résidus, le rapport $\hat{R}_{k,5/2}$ est un estimateur biaisé de $R_{k,5/2} = \{E\ (Y_{5K.})\ /\ E\ (Y_{2k.})\}$ mais nous avons constaté que le biais a un effet négligeable sur l'erreur quadratique moyenne (EQM) estimée des estimateurs et par conséquent, il n'a pas été pris en compte au moment de la construction de l'indice.

Une fois les indices par case calculés, ils sont agrégés à un niveau supérieur en fonction des poids des coûts appropriés obtenus de l'enquête la plus récente sur les dépenses des familles. Les indices mensuels sont calculés par interpolation en utilisant comme données repères les changements correspondants de l'indice des prix des "entrées dans la construction résidentielle ". Les indices mensuels sont ensuite intégrés à l'IPC.

En raison de l'enregistrement tardif de certaines ventes et des retards administratifs de traitement, les données relatives à un mois donné peuvent devenir disponibles jusqu'à trois mois plus tard. Ces données sont prises en compte au moment de la révision dont l'IPL fait l'objet après trois mois, laquelle coïncide avec le calcul du nouvel IPL. Toutefois, l'IPL révisé, quoique plus stable, est d'utilisation restreinte.

ANALYSE: La méthode que nous venons de décrire comporte des faiblesses évidentes. L'hypothèse voulant que les effets marginaux des VMQ demeurent fixes tout au long d'une période de six mois et que les changements de prix soient uniquement pris en compte dans la fonction de temps va à l'encontre de la plupart des études sur les indices réalisées à ce jour et elle est avant tout une approximation. Autrement dit, dans l'équation

L'article suit la présentation suivante: dans la prochaine section, nous passons en revue la méthode de la régression hédonique, laquelle sert à l'ajustement des variations de qualité, et nous décrivons son application en Israël en faisant ressortir les problèmes que pose son utilisation. Dans la section 3, nous définissons le modèle proposé et analysons ses propriétés. L'estimation des paramètres du modèle est le sujet de la section 4. Dans la section 5, nous proposons une version modifiée du modèle afin d'assurer sa robustesse et de contrôler son rendement en période de montée rapide de l'inflation. La section 6 présente des résultats empiriques qui illustrent les caractéristiques importantes du modèle. Enfin, dans la section 7, nous concluons l'article en donnant un aperçu des travaux à faire en vue d'une analyse plus poussée de la question.

Comme cet article vise avant tout à fournir une vue d'ensemble du modèle, nous nous sommes efforcés de restreindre le plus possible les détails techniques. Les opérations mathématiques manquantes peuvent être obtenues des auteurs.

2. AJUSTEMENT POUR TENIR COMPTE DES CHANGEMENTS DE QUALITÉ À PARTIR DE RÉGRESSIONS HEDONIQUES

La méthode courante d'ajustement pour tenir compte des changements de qualité consiste à recourir à la régression "hédonique" découlant des travaux de Court (1939), Stone (1956) et Adelman et Griliches (1961). (La première et la troisième de ces études portent sur le calcul d'indices de prix d'automobiles et la deuxième a trait aux indices de prix pour l'établissement des comptes nationaux.)

La méthode de la régression hédonique comporte deux variantes. Dans la première, les prix de vente correspondant à une période donnée sont soumis à une régression en fonction de variables de mesure de la qualité (VMQ). A l'aide des coefficients estimés, le "prix de vente moyen" est estimé pour chacune des périodes à l'étude en calculant les valeurs de régression après ajustement qui correspondent à des valeurs fixes "moyennes" des VMQ. Le calcul des rapports entre ces moyennes produit les indices souhaités. Dans la seconde variante, les prix de vente de plusieurs périodes font l'objet d'une régression en fonction des VMQ et de variables fictives dont les coefficients représentent des estimations du changement de prix pur. (On suppose que les coefficients de régression des autres variables sont fixes pour l'ensemble des périodes à l'étude.)

Le raisonnement sous-jacent dans les deux cas est que la variation des prix de vente est attribuable "en grande partie" à un nombre relativement restreint de VMQ (appelées caractéristiques dans le contexte de la régression hédonique) et que les autres aspects, non pris en compte, des variations de qualité n'ont aucune corrélation avec ceux qui sont inclus. Dans le cas de la première variante, les coefficients de régression peuvent varier avec le temps tandis que dans la seconde, on suppose que les poids des VMQ sont fixes, en d'autres termes que tout changement dans les prix moyens entre les périodes successives à l'étude est pris en compte dans les coefficients des variables de temps fictives. En supposant que chacune des équations de régression utilisées pour la première variante comprennent les ordonnées à l'origine, on se rend compte que le modèle combiné appliqué à l'ensemble des périodes à l'étude comprend le modèle utilisé pour la seconde variante en tant que cas particulier. Les aspects théoriques associés à l'utilisation de ces deux méthodes sont analysés dans Griliches (1971). (Voir aussi l'analyse présentée à la fin de cette section.)

En Israël, le Central Bureau of Statistics (CBS) a adopté une version modifiée de la seconde variante pour le calcul des IPL. Trois VMQ sont utilisées dans la régression: la superficie (en mètres carrés), l'âge (en années) et le district (défini par une ou deux variables fictives selon la taille de la ville).

Les calculs se font en trois étapes:

Etape 1 Pour chaque case définie selon la ville et le nombre de pièces, lorsque le nombre des données est suffisant, un modèle de régression multiplicatif est estimé tous les trois mois à l'aide des données disponibles pour la période de six mois la plus récente. L'équation de régression a la forme

$$\text{Log } Y_{tkj} = \alpha_o + \alpha_{k1} \log F_{tkj} + \alpha_{k2} \log A_{tkj} + \alpha_{k3} D^{(1)}_{tkj} + \alpha_{k4} D^{(2)}_{tkj} + g_k(t) + \varepsilon_{tkj} \qquad \textbf{(2.1)}$$

où Y_{tkj} est le prix de la j-ième vente dans la case k au cours du mois t, F_{tkj}, A_{tkj}, $D^{(1)}_{tkj}$ and $D^{(2)}_{tkj}$ représentent la superficie, l'âge et les deux variables "indicateurs" du district (une seule de ces variables est utilisée dans le cas des petites villes) et où ε_{tkj} est un résidu aléatoire auquel on associe une variance constante $\sigma^2_k = E(\varepsilon^2_{tkj})$.

La fonction de temps $g^{(t)}_k$ est linéaire par morceaux et est définie pour t=1 ... 6 de la manière suivante (t=6 représente le mois le plus récent pour lequel on dispose de données) certains vêtements sont saisonniers et ne sont pas disponibles pour toutes les périodes de référence. Parmi les biens durables, de nouveaux modèles ne cessent d'être créés et leur qualité diffère parfois de celle des modèles sortis antérieurement. C'est le problème auquel on fait face chaque fois au moment de calculer les indices de prix annuels des véhicules automobiles.

Dans le calcul des indices de prix du logement (IPL), les variations de qualité d'une période de référence à une autre sont attribuables au fait que les ventes enregistrées au cours de deux périodes différentes quelconques ne sont soumises à aucun contrôle et qu'elles ont habituellement trait à des genres d'habitation différents. Ce problème est particulièrement important en Israël, d'où viennent nos données empiriques, étant donné que l'IPL global est une moyenne pondérée des IPL calculés pour de petites cases classées selon l'unité géographique (ville) et la taille de la maison (nombre de pièces). Dans des intervalles aussi courts qu'un mois, le nombre réel de ventes effectuées et enregistrées peut être très faible dans nombre de ces cases, ce qui est source d'écarts importants sur le plan de la qualité.

A titre d'exemple, les graphiques ci-dessous montrent les moyennes mensuelles d'âge et de superficie d'appartements de deux pièces situés dans la ville de Jérusalem et vendus entre les mois de juillet 1987 et juin 1989. Le nombre de ventes sur lesquelles ces moyennes sont fondées varie entre 5 et 69. (Le nombre des ventes est particulièrement faible pour les trois derniers mois parce que la plupart d'entre elles sont habituellement enregistrées dans les trois mois qui suivent les premiers calculs et la publication de l'IPL.)

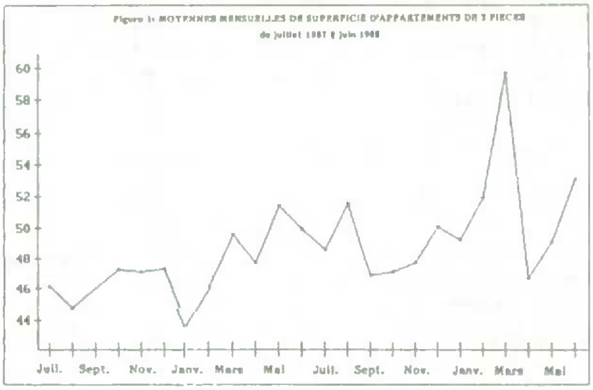


Figure 1: MOYENNES MENSUELLES DE SUPERFICIE D'APPARTEMENTS DE 2 PIECES de juillet 1987 à juin 1989

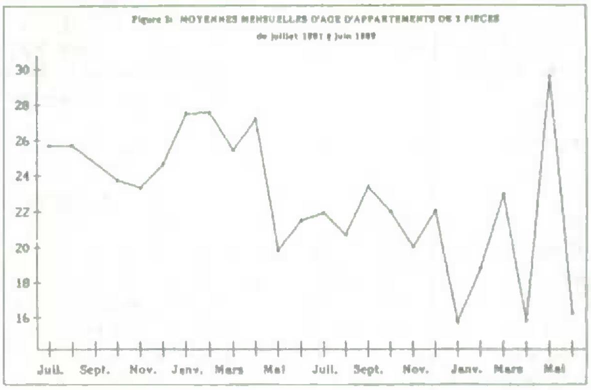


Figure 2: MOYENNES MENSUELLES D'AGE D'APPARTEMENTS DE 2 PIECES de juillet 1987 à juin 1989

Le problème des variations de qualité aux fins de calcul d'indices de prix comporte plusieurs facettes et il a été analysé par de nombreux auteurs. Citons par exemple les ouvrages de Hofsten (1952) et de Griliches (1971). (Les deux études portent le même titre - "Price Indexes and Quality Change".) Cependant, la plupart des études sur le sujet mettent l'accent sur le calcul d'indices généraux de prix de biens durables et visent donc principalement à trouver des façons de tenir compte des améliorations techniques et de l'ajout de nouvelles caractéristiques plutôt que des changements de qualité causés par la petite taille des échantillons. Comme le fait remarquer Griliches, "la plupart des chercheurs dans ce domaine, y compris moi-même, ont essayé d'obtenir le plus grand échantillon transversal que possible pour toute année donnée, sans trop se préoccuper de la comparabilité générale de deux quelconques de ces échantillons" (Griliches, 1971, p.7).

Dans le présent article, nous nous intéressons à cet aspect précis du problème des variations de qualité, c'est-à-dire celles causées par l'utilisation d'échantillons de petite taille qui ne sont pas soumis à un contrôle. Nous mettons l'accent sur le calcul d'indices de prix du logement fondés sur des prix de vente réels. En Israël (comme dans de nombreux autres pays), l'IPL est une composante importante de l'IPC avec un poids d'environ 15%. L'IPL est aussi un indicateur économique clé et il sert à des fins de couplage de contrats de construction et de location de maisons.

Soulignons que l'utilisation des prix de vente réels (souvent désignée par la méthode fondée sur les achats de maisons dans les ouvrages traitant de la question) n'est qu'une possibilité parmi plusieurs autres méthodes de calcul de l'IPL. De fait, il existe au moins quatre méthodes différentes et les méthodes utilisées changent avec les pays et avec les années. Ainsi, le Bureau of Labour Statistics des Etats-Unis a utilsé jusqu'en 1983 la méthode fondée sur les achats de maisons au moment où il décidait d'adopter la méthode dite des équivalents de location tandis qu'en Nouvelle-Zélande on passait plutôt de la méthode des équivalents de location à la méthode des achats de maison. Castles (1987) fournit une excellente revue des diverses méthodes et résume les manières de procéder dans plus de 130 pays.

Bien que nous envisagions le problème dans le contexte des indices de prix du logement, la méthode décrite dans le présent article peut être appliquée, avec certaines modifications, à d'autres indices de prix de même nature, par exemple le calcul d'indices de prix de véhicules d'occasion. De plus, le modèle que nous utilisons est un modèle de régression avec coefficients stochastiques qui peuvent varier de façon tant transversale que longitudinale. Un tel modèle a des applications très variées dans des études statistiques et économétriques.

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MODÈLE DE SÉRIE CHRONOLOGIQUE AJUSTÉ POUR TENIR COMPTE DES VARIATIONS DE QUALITÉ ET SERVANT À L'ESTIMATION DES INDICES DE PRIX DU LOGEMENT

D. Pfeffermann, L. Burck et S. Ben-Tuvia[1]

RÉSUMÉ

L'estimation des indices de prix du logement se fonde sur les prix de vente de maisons enregistrés au cours de périodes successives. Cependant, il n'est pas possible d'exercer un contrôle sur l'enregistrement des ventes de maisons et celles-ci concernent des maisons de qualité différente vendues au cours de périodes différentes. Une méthode courante d'ajustement pour tenir compte des variations de qualité (comme celle utilisée dans le calcul des indices de prix des automobiles) consiste à régresser les prix de vente en fonction de diverses variables de mesure de la qualité. Toutefois, un indice de prix du logement étant calculé pour chaque case, dans de nombreux cas, le nombre des ventes est minime ou même nul dans la case au moment de la détermination de l'indice. Afin de remédier à ce problème, nous proposons l'utilisation d'un modèle linéaire dynamique qui tient compte des rapports chronologiques entre les coefficients de régression des cases et permet d'établir des corrélations actuelles entre coefficients associés à des cases voisines. Nous proposons également des modifications au modèle afin d'assurer sa robustesse et de contrôler son rendement en période de montée rapide de l'inflation. Nous présentons enfin des résultats empiriques qui comparent le rendement du modèle à celui de modèles fondés sur des coefficients de régression fixes en nous servant à cette fin de données sur les prix de maisons vendues dans la ville de Jérusalem entre les années 1982 et 1989.

MOTS CLES: Régression hédonique, indice de Laspeyres, prévision robuste, modèle d'espace d'états.

1. INTRODUCTION

L'indice des prix à la consommation (IPC) est une des séries économiques les plus importantes et les plus utilisées. Il constitue un indicateur clé du développement économique et sert souvent de base de calcul dans les négociations salariales et pour certaines opérations des marchés financiers. Une autre utilisation importante de l'IPC est comme déflateur pour convertir des séries statistiques exprimées en prix courants en séries exprimées en prix constants pour une période donnée.

Idéalement, l'IPC vise à mesurer l'effet des changements de prix sur le budget dont doivent disposer les consommateurs s'ils veulent maintenir un certain niveau de consommation. En pratique, l'indice permet de mesurer la variation en pourcentage dans le temps des sommes devant être consacrées à la consommation d'un "panier" fixe de biens et services. Les éléments qui composent le panier ainsi que leurs poids relatifs sont déterminés périodiquement en fonction des résultats d'une enquête sur les dépenses des familles, de façon à assurer que le panier représente la consommation moyenne de la population à laquelle l'indice s'applique.

Nous nous limiterons dans cette étude à l'analyse de l'indice de Laspeyres, l'indice le plus courant. En supposant que P_{ko} et Q_{ko} représentent le prix et la quantité d'un élément k au cours d'une période de base et que P_{kt} est le prix correspondant du même élément au cours de la période t, l'indice de Laspeyres est défini par

$$L_t = \sum_k P_{kt} Q_{ko} \Big/ \sum_k P_{ko} Q_{ko} = \sum_k \frac{P_{kt}}{P_{ko}} W_k \tag{1.1}$$

où la somme est celle de tous les éléments compris dans le panier et où $W_k = P_{ko} Q_{ko} / \sum_k P_{ko} Q_{ko}$. Ecrit sous cette forme, l'indice peut être considéré comme une moyenne pondérée des indices de prix $R_{kt} = (P_{kt}/P_{ko})$ des biens et services inclus dans le panier, où les poids représentent les dépenses relatives concernant les mêmes éléments effectuées au cours de la période de base. L'élément k peut être en soi un agrégat d'un certain nombre de sous-éléments; dans ce cas, l'indice R_{kt} est calculé à nouveau en tant qu'indice de Laspeyres des sous-éléments qui composent l'élément k. La méthode est habituellement appliquée à plusieurs niveaux d'agrégation, selon le bien ou le service à l'étude.

Afin d'assurer que l'indice ne reflète que les variations de prix des biens et services et aucun autre changement, il est essentiel que les prix enregistrés au cours de périodes successives s'appliquent à des éléments identiques ou équivalents. Cependant, ce genre d'exigence pose souvent des problèmes. En effet, certains aliments et

[1] D. Pfeffermann, Hebrew University, Jerusalem, Israel 91905; L. Burck, Central Bureau of Statistics, Jerusalem, Israel 91130; S. Ben-Tuvia, Central Bureau of Statistics, Jerusalem, Israel 91130.

SECTION 2

L'ANALYSE DES SÉRIES CHRONOLOGIQUES EN PRÉSENCE D'ERREURS D'ENQUÊTES

l'utilisation dans les modèles à moyennes fixes, comme (7), ou les modèles à moyennes aléatoires qui utilisent une méthode d'estimation en deux étapes, comme celui décrit dans la sous-section 3.3. Les modèles d'espace d'états peuvent être appliqués facilement dans les circonstances pour évaluer la fonction de vraisemblance marginale. Par ailleurs, lorsque le nombre de paramètres de nuisance est peu élevé, comme dans le modèle à moyennes aléatoires défini en (22), il est préférable de recourir à la fonction de vraisemblance intégrale.

REMERCIEMENTS

Cette étude a été rendue possible grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

Bellhouse, D.R. (1978). Marginai Likelihoods for distributed lag models. *Statist. Hefte* **19**: 2-14.

Bellhouse, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *J. Statist. Plan. Inf.* **21**: 69-74.

Bellhouse, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, to appear.

Binder, D.A. and Hidiroglou, M.A. (1988). Sampling in time. In: *Handbook of Statistics, Volume 6 (Sampling)*, P.R. Krishnaiah and C.R. Rao (eds.). Amsterdam: North-Holland, pp. 187-211.

Binder, D.A. and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology* **15**: 29-45.

Blight, B.J.N. and Scott, A.J. (1973). A stochastic model for repeated surveys. *J. Roy. Statist. Soc. (B)* **35**: 61-68.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. (B)* **49**: 1-39.

Cruddas, A.M., Reid, N., and Cox, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76**: 231-237.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Ann. Math. Statist.* **38**: 1456-1465.

Harvey, A.C. and Phillips, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika* **66**: 49-58.

Kalbfleisch, J.D. and Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. (B)* **32**: 175-208.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *J. Roy. Statist. Soc. (B)* **12**: 241-255.

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.* **20**: 335-375.

Roberts, G., Rao, J.N.K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**: 1-12.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *J. Roy. Statist. Soc. (B)* **51**: 15-27.

approximativement. Nous pouvons établir un modèle en supposant, pour G_r, un processus ARMA comme celui ci-dessous

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} = \phi(\bar{y}_{t-1,r} - \mu_{t-1})/(\text{deff}_{t-1,r})^{1/2} + \epsilon_t, \qquad (21)$$

où ϵ_t a une variance constante. Cela cadre bien avec le modèle (1), où le vecteur d'observations y renferme des données de la forme $\bar{y}_{t,r}/(\text{deff}_{t,r})^{1/2}$, β est $(\mu_1, \mu_2, ..., \mu_k)^T$, et X renferme des éléments de la forme $1/(\text{deff}_{t-1,r})^{1/2}$. La fonction de vraisemblance marginale, considérée en l'occurrence comme un cas particulier des équations (5) ou (6), peut être évaluée à l'aide du modèle d'espace d'états proposé par Harvey et Phillips (1979) et dont nous avons fait mention dans la section 2. Compte tenu du modèle défini ci-dessus (équ. 20 et 21), il est souhaitable de recourir à l'estimation fondée sur la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative. La valeur estimée de ϕ repose en l'occurrence sur la variation entre les estimations élémentaires dans chaque groupe de renouvellement, la variance de ces estimations n'étant pas connue. Comme un groupe de renouvellement passe relativement peu de temps dans l'échantillon, il y a de fortes chances que les estimateurs du maximum de vraisemblance soient biaisés et non-convergents.

Si nous combinons le modèle (21) avec le modèle (10) par exemple, nous pouvons alors utiliser la méthode en deux étapes décrite dans la sous-section 3.3 pour estimer le paramètre autorégressif en (10).

Pour ce qui a trait au second scénario, supposons que nous connaissons les valeurs estimées de la moyenne, $\bar{y}_t$, pour chaque répétition $t = 1, ..., k$. Supposons aussi que la matrice, S, des variances-covariances des estimations est connue. Il est alors possible de déduire de l'équation (6) une pseudo fonction de vraisemblance marginale. Comme dans Binder et Dick (1989) notamment, les $\bar{y}_t$ peuvent être définies par le modèle

$$\bar{y}_t = \mu_t + e_t, \qquad (22)$$

où e_t est l'erreur d'enquête à la t-ième répétition, la matrice des variances-covariances estimée étant représentée par S. Les moyennes pour chaque répétition (μ_t pour la t-ième répétition) suivent un processus ARMA. Comme il s'agit là d'un cas particulier du modèle de régression avec coefficients aléatoires, il est possible de déduire de l'équation (6) la fonction de vraisemblance marginale appropriée. Puisque S est connue, nous pouvons obtenir facilement une estimation d'Ω, la matrice de corrélation des erreurs d'enquête. Nous pouvons aussi obtenir une valeur estimée de $\kappa = \sigma^2/\gamma^2$. Les hypothèses qui sous-tendent la fonction de vraisemblance marginale définie en (6) nous obligent à supposer que e_t dans l'équation (22) est une variable aléatoire stationnaire. Par conséquent, la moyenne des éléments diagonaux de S donne une valeur estimée de σ^2. Si γ^2 est la variance des moyennes μ, alors la variation entre les $\bar{y}_t$, $t = 1, ..., k$, donne une valeur estimée de $\sigma^2 + \gamma^2$. De ces deux valeurs estimées, nous pouvons déduire une valeur estimée pour κ. Suivant le modèle (22), X dans l'équation (6) est la matrice unité $k \times k$ tandis que W est un vecteur colonne $k \times 1$ formé de uns. Nous pouvons donc déterminer la pseudo fonction de vraisemblance marginale pour Γ (pseudo car κ et Ω ont été remplacés par leurs estimateurs) en prenant l'équation (6) et en effectuant les substitutions appropriées. Si k, le nombre de répétitions, est relativement élevé par rapport au nombre de paramètres dans Γ, les estimateurs fondés sur la fonction marginale et la fonction conditionnelle approximative devraient être semblables à l'estimateur du maximum de vraisemblance. Au point de vue du calcul, il semble que la fonction de vraisemblance intégrale qui utilise les modèles d'espace d'états décrits par Binder et Dick (1989, section 3) soit la plus simple à appliquer dans les circonstances.

5. ANALYSE

La fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative sont des méthodes d'estimation qui peuvent être utilisées dans diverses formes d'échantillonnage répété. Comme les fonctions de vraisemblance marginales sont beaucoup plus efficaces que les estimateurs du maximum de vraisemblance lorsque le nombre de paramètres de nuisance est élevé, nous pouvons en recommander

La seconde façon d'intégrer des hypothèses comme l'équation (18) à la méthode d'estimation comporte deux étapes. Suivant le modèle des paramètres de régression fixes, la FM ou la FCA se présentent sous une forme très élémentaire, qui est exprimée par les équations (4) et (5). En ce qui concerne les enquêtes répétées avec modèle autorégressif du premier degré et échantillonnage aléatoire simple, la FM et la FCA sont définies par l'équation (17). De plus, on peut déterminer facilement la valeur de la FM pour n'importe quelle valeur donnée des paramètres d'Ω par une application directe du modèle d'espace d'états de Harvey et Phillips (1979). Lorsqu'on utilise le modèle des coefficients aléatoires (par exemple le modèle de Blight et Scott (1973) dans les enquêtes répétées), les fonctions de vraisemblance (complète, marginale ou conditionnelle approximative) et les modèles d'espace d'états pertinents deviennent beaucoup plus compliqués. En outre, il est difficile de définir des modèles comme (18) ou un processus de degré supérieur. Pour plus de simplicité, quitte à sacrifier un peu d'efficacité, nous proposons ci-dessous une méthode pour estimer les paramètres d'Ω. On estime tout d'abord les paramètres d'Ω à l'aide de la fonction de vraisemblance marginale ou de la fonction de vraisemblance conditionnelle étant donné les moyennes des répétitions $\mu_1, \ldots, \mu_k$ (ou étant donné β dans le cas d'une régression). Le nombre de paramètres du modèle s'accroît avec le nombre de répétitions k. Lorsque les séries d'observations sur des unités individuelles sont relativement courtes, comme c'est le cas dans les enquêtes répétées, les estimateurs du maximum de vraisemblance des paramètres d'Ω peuvent être biaisés et non-convergents. Il est toutefois possible de remédier à cette difficulté, comme le montrent empiriquement Cruddas et coll. (1989) pour un processus autorégressif du premier degré, en utilisant la FM ou la FCA pour estimer les paramètres de corrélation. Une fois que les paramètres d'Ω ont été estimés, on peut déterminer les valeurs estimées $\hat{\mu}_1, \ldots, \hat{\mu}_k$ de $\mu_1, \ldots, \mu_k$ à l'aide des méthodes décrites par Harvey et Phillips (1979). Si nous prenons par exemple le modèle composé (10) et (18), le processus en (18) étant remplacé par un processus ARMA général, la matrice des variances-covariances de $\hat{\mu}_1, \ldots, \hat{\mu}_k$ est définie par l'expression $\sigma^2 G^{-1} + \gamma^2 \Gamma$. Si $\sigma^2 G^{-1}$ est faible par rapport à $\gamma^2 \Gamma$, ce qui peut être le cas lorsque la taille des échantillons pour les estimations élémentaires d'enquête est élevée, alors on peut se servir des valeurs estimées $\hat{\mu}_1, \ldots, \hat{\mu}_k$ comme données pour définir le processus et estimer les paramètres de Γ et ce, sans perte notable d'efficacité. On peut ensuite établir des estimations révisées de μ_t au moyen du processus estimé.

4. ENQUÊTES À PLAN DE SONDAGE COMPLEXE

Il y a plusieurs façons d'analyser des données de séries chronologiques tirées d'enquêtes à plan de sondage complexe. Chaque méthode que l'on peut proposer dépendra des données d'échantillon qui auront pu être recueillies.

Si, par exemple, on dispose de micro-données, il est possible de calculer pour chaque groupe de renouvellement la matrice, fondée sur le plan de sondage complexe, des variances- covariances des estimations élémentaires. Lorsque les moyennes $\mu_1, \ldots, \mu_k$ sont considérées comme fixes, une pseudo fonction de vraisemblance marginale est définie par les équations (4) et (9), où $\hat{x}_r$ et S_r sont remplacés par leurs équivalents pour les enquêtes à plan de sondage complexe. C'est la méthode qu'utilisent par exemple Roberts, Rao et Kumar (1987) dans l'analyse de régression logistique pour plans de sondage complexes: déterminer une fonction de vraisemblance ou un ensemble d'équations de vraisemblance et remplacer les paramètres statistiques habituels par leurs équivalents pour les enquêtes à plan de sondage complexe. Pour ce qui a trait aux moyennes de modèles aléatoires, on peut recourir à l'analyse de moyennes fixes dans la première étape de la méthode d'estimation décrite dans la sous-section 3.3. On peut aussi déterminer la fonction de vraisemblance marginale suivant le modèle des moyennes aléatoires, par exemple en déterminant tout d'abord la fonction de vraisemblance définie en (19), puis la fonction marginale correspondante. On remplace ensuite les paramètres statistiques de cette fonction marginale par leurs équivalents dans les plans de sondage complexes pour obtenir une pseudo fonction de vraisemblance marginale.

Or, il arrive rarement que l'on dispose de micro-données. La méthode d'estimation dépend alors des données disponibles. Nous envisageons ici deux scénarios mais beaucoup d'autres sont possibles. Dans le premier scénario, les covariances ou les corrélations d'échantillon ne sont pas connues alors que dans le second, elles le sont.

Supposons que nous connaissons seulement les estimations élémentaires et les effets du plan correspondants. Soit $\bar{y}_{t,r}$ l'estimation tirée du groupe de renouvellement G_r à la t-ième répétition et fondée sur un échantillon de taille m_r. Soit $\mathrm{deff}_{t,r}$ l'effet du plan qui correspond à $\bar{y}_{t,r}$. Si σ^2/m_r est la variance de $\bar{y}_{t,r}$ suivant un échantillonnage aléatoire simple, alors, en vertu du théorème limite central,

$$(\bar{y}_{t,r} - \mu_t)/(\mathrm{deff}_{t,r})^{1/2} \sim N(0,\sigma^2/m_r) \tag{20}$$

$$g_{tt} = \pi_t n_t + (1 - \pi_t) n_t (1 - \Phi^2) + \pi_{t+1} n_{t+1} \Phi^2, \text{ pour } t = 1, \ldots, k \tag{14}$$

et

$$g_{t,t+1} = -\pi_{t+1} n_{t+1} \Phi, \text{ pour } t = 1, \ldots, k-1, \tag{15}$$

où $\pi_1 = \pi_{k+1} = 0$. Les éléments de z sont définis

$$z_t = \pi_t n_t (\bar{y}'_t - \Phi \bar{x}'_{t-1}) + (1 - \pi_t) n_t \bar{y}''_t (1 - \Phi^2) - \pi_{t+1} n_{t+1} (\bar{y}'_{t+1} - \Phi \bar{x}'_t), \tag{16}$$

pour $t = 1, \ldots, k$, où $\pi_1 = \pi_{k+1} = 0$ et $\bar{y}''_1 = \bar{y}_1$. Le vecteur des moyennes estimées $\hat{\mu}$ est non biaisé pour μ selon le modèle (10) et la matrice des variances-covariances correspondante est $\sigma^2 G^{-1}$. Ainsi, d'après les équations (4) ou (5), la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative pour Φ est

$$L_M(\phi) = \frac{(1 - \delta^2)^{d/2}}{\{A(\hat{\mu},\phi) + B(\phi)\}^{(m-k)/2} |G|^{1/2}}. \tag{17}$$

3.3 Moyennes de modèle aléatoires

Dans les deux sous-sections précédentes, nous n'avons pas considéré le fait qu'il peut exister un rapport entre les moyennes établies pour chaque répétition. Ces moyennes représentent l'information la plus précieuse et beaucoup d'information pourrait être perdue si l'on ne tenait pas compte de la relation entre les moyennes d'une période à l'autre. Blight et Scott (1973), par exemple, soulignent que les moyennes établies lors d'enquêtes répétées sont souvent corrélées et supposent, en plus du modèle (10), que

$$\mu_t - \xi = \phi(\mu_t - \xi) + u_t, \tag{18}$$

où $u_t \sim N(0, \gamma^2)$ et où les u sont mutuellement indépendants. Il existe au moins deux façons d'intégrer des hypothèses comme l'équation (18) dans la méthode d'estimation.

La première est d'utiliser l'approche intégrale de la vraisemblance. Selon le modèle défini par les équations (10) et (18), la fonction de vraisemblance logarithmique pour les données devient

$$L(\xi,\gamma^2,\Theta,\phi,\kappa) = -m \ln\gamma + (k/2) \ln\kappa + (d/2) \ln(1 - \Phi^2) + (1/2) \ln(1 - \phi^2)$$

$$- \{A(\hat{\mu},\Phi) + B(\Phi) + C(\Phi,\phi,\kappa) - 2(\mu\kappa) D(\Phi,\phi,\kappa) + (\mu^2\kappa) E(\Phi,\phi,\kappa)\}/2\gamma^2). \tag{19}$$

Dans l'équation ci-dessus, $C(\Phi,\phi,\kappa) = z^T(G^{-1} - F^{-1})z$, $D(\Phi,\phi,\kappa) = (1 - \phi) \nu^T F^{-1} z$ et $E(\Phi,\phi,\kappa) = k -2(k-1)\phi + (k-2)\phi^2 + \kappa(1 - \phi^2 \nu^T F^{-1} \nu$, où le vecteur 1 x k $\nu^T = (1, 1-\phi, 1-\phi, \ldots, 1-\phi, 1)$, la matrice Gest définie par (14) et (15) et z est défini par (16). La matrice F dans l'équation ci-dessus est une matrice symétrique k x k en bande largeur 3, dont les éléments diagonaux sont $g_{tt} + \kappa(1 + \phi^2)$ pour $t = 2, \ldots, k-1$ et $g_{tt} + \kappa$ pour $t = 1$ ou k, et les éléments non diagonaux non-nuls sont $g_{t,t+1} + \phi\kappa$ pour $t = 1, \ldots, k-1$. En posant les dérivées de (19) par rapport à ξ et à γ égales à 0, on peut déterminer facilement l'estimateur du maximum de vraisemblance de ces paramètres, étant donné Φ, ϕ et κ. De plus, en calculant la variance de $\hat{\xi}$ selon (10) et (18) et en utilisant l'équation (6), on peut exprimer facilement la fonction de vraisemblance marginale et la fonction conditionnelle approximative, $L_M(\Phi,\phi,\kappa)$, même si elle est une fonction compliquée de Φ, ϕ et κ. Comme le nombre total de paramètres est peu élevé, l'estimateur du maximum de vraisemblance et l'estimateur marginal du maximum de vraisemblance sont tous deux convergents et asymptotiquement non biaisés et auront probablement des valeurs comparables. Bien qu'il puisse s'agir d'expressions complexes, les fonctions de vraisemblance exactes peuvent être déterminées en remplaçant (10) et (18) par un modèle autorégressif de moyennes mobiles stationnaire général. De même, il est possible de calculer les fonctions de vraisemblance marginales et conditionnelle approximative correspondantes.

dans la sous-section 3.1, le vecteur des paramètres de régression β est $(\mu_1, \ldots, \mu_k)^T$. Lorsque le vecteur de données **y** contient pour chaque unité les observations groupées selon les passages où cette unité a été échantillonnée, selon l'échantillonnage avec renouvellement de la sous-section 3.1, on peut exprimer la matrice de corrélation Ω, qui est désormais une fonction de Φ, comme une somme directe de matrices qui sont l'une et l'autre les matrices de corrélation d'un processus autorégressif du premier degré.

Nous reprenons la notation utilisée par Patterson (1950) pour désigner des tailles d'échantillon, des moyennes ainsi que des sommes des carrés et des produits (centrées sur leur moyenne appropriée) pour le passage t:

$\overset{*}{\pi}_t$ = la proportion d'unités échantillonnées à la t-ième répétition, qui étaient aussi présentes dans l'échantillon à la répétition précédente (t-1);

n_t = le nombre d'unités échantillonnées à la t-ième répétition t;

$\bar{y}'_t$ = la moyenne pour les unités échantillonnées à la t-ième répétition, qui étaient aussi présentes dans l'échantillon à la répétition précédente (t-1);

$\bar{y}''_t$ = la moyenne pour les unités échantillonnées à la t-ième répétition et qui n'étaient pas présentes dans l'échantillon à la répétition précédente (t-1);

$\bar{y}_t$ = la moyenne pour toutes les unités échantillonnées à la t-ième répétition;

$\bar{x}'_t$ = la moyenne pour les unités échantillonnées à la t-ième répétition, qui sont aussi présentes dans l'échantillon à la répétition suivante (t+1);

syy'_t = la somme des carrés pour les unités échantillonnées à la t-ième répétition, qui étaient aussi présentes dans l'échantillon à la répétition précédente (t-1);

syy''_t = la somme des carrés pour les unités échantillonnées à la t-ième répétition et qui n'étaient pas présentes dans l'échantillon à la répétition précédente (t-1);

sxx'_t = la somme des carrés pour les unités échantillonnées à la t-ième répétition, qui sont aussi présentes dans l'échantillon à la répétition suivante (t+1);

sxy'_t = la somme des carrés pour toutes les unités échantillonnées à la t-ième répétition;

syy_t = la somme des produits pour les observations relatives aux unités échantillonnées à la t-ième répétition, qui étaient aussi présentes dans l'échantillon à la répétition précédente t-1.

Suivant le cas particulier du modèle (10), et après de nombreuses transformations algébriques, nous pouvons montrer que, lorsqu'on fait la sommation de l'expression (8) pour tous les groupes de renouvellement r, la fonction de vraisemblance logarithmique pour les données se ramène à

$$l(\mu_1, \ldots, \mu_k, \sigma^2, \Phi) = -m \ln\sigma + (d/2)\ln(1-\Phi^2) - \{A(\mu,\Phi) + B(\Phi)\}/(2\sigma^2), \tag{11}$$

où d est le nombre d'unités échantillonnées distinctes (c'est-à-dire abstraction faite du nombre de fois qu'une unité est échantillonnée) et m est la taille de l'échantillon global $(n_1 + \ldots + n_k)$. De plus, dans l'équation (11),

$$A(\mu,\Phi) = (1-\Phi^2)n_1(\bar{y}_1 - \mu_1)^2 + \sum_{t=2}^{k} [\pi_t n_t\{\bar{y}'_t - \mu_t - \Phi(\bar{x}'_{t-1} - \mu_{t-1})\}^2 + (1 - \pi_\tau)n_t(1 - \Phi^2)(\bar{y}''_t - \mu_t)^2] \tag{12}$$

et

$$B(\Phi) = (1-\Phi^2)\ syy_1 + \sum_{t=2}^{k} \{\Phi^2\ sxx'_{t-1} - 2\Phi\ sxy'_t + syy'_t + (1-\Phi^2)\ syy''_t\}. \tag{13}$$

Pour n'importe quelle valeur donnée de Φ, les estimateurs du maximum de vraisemblance sont $\hat{\mu} = G^{-1}z$ et $\hat{\sigma}^2 = \{A(\hat{\mu}, \Phi) + B(\Phi)\}/m$, où $A(\hat{\mu},\Phi)$ est défini par l'équation (12), μ étant remplacé par son estimateur du maximum de vraisemblance, et où G est une matrice symétrique k x k en bande de largeur 3 et z est un vecteur k x 1. Les éléments non-nuls de G sont définis

Nous reprenons la notation de Bellhouse (1989) pour décrire le plan d'échantillonnage. À chaque tenue de l'enquête, on échantillonne c groupes de renouvellement. Le groupe de renouvellement r, désigné par G_r, contient m_r unités, $r = 1, 2, \ldots, k + c - 1$. Pour la t-ième répétition, l'échantillon comprend les unités des groupes G_t, G_{t+1}, ..., G_{t+c-1}, de sorte que sa taille n_t est égale à $m_t + m_{t+1} + \ldots + m_{t+c-1}$. Chaque groupe de renouvellement est choisi selon un échantillonnage aléatoire simple sans remise parmi des unités de la population qui n'ont pas été prélevées auparavant. Pour les k répétitions considérées globalement, la taille de l'échantillon est $m = n_1 + n_2 + \ldots + n_k$.

Supposons que G_r est choisi la première fois à la u-ième répétition et la dernière fois à la v-ième répétition, u étant égal à 1 ou à r et v correspondant à $r + c - 1$ où à k. Le nombre total de répétitions où une unité de G_r est incluse dans l'échantillon est $b = v + 1 - u$. Soient $\bar{y}_{u,r}, \ldots, \bar{y}_{v,r}$ les moyennes d'échantillon ou les estimations élémentaires pour G_r aux répétitions u, $u + 1$, ..., $v - 1$, v respectivement. Alors, suivant le modèle (7), la contribution de G_r à la fonction de vraisemblance logarithmique définie en (2) est

$$-\{bn_r \ln\sigma + (n_r/2)\ln(|\Omega_r|) + [n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1)\,\mathrm{tr}(\Omega_r^{-1} S_r)]/(2\sigma\}, \tag{8}$$

où x_r^T est le vecteur $1 \times b$ $(\bar{y}_{u,r} - \mu_u, \bar{y}_{u+1,r} - \mu_{u+1}, \ldots, \bar{y}_{v-1,r} - \mu_{v-1}, \bar{y}_{v,r} - \mu_v)$, S_r est la matrice $b \times b$ des sommes des carrés et des produits pour les observations relatives au groupe de renouvellement et où Ω_r est la matrice de corrélation $b \times b$ des observations relatives à une unité du groupe de renouvellement. En vertu de l'hypothèse d'indépendance, on obtient la fonction de vraisemblance logarithmique complète en faisant la sommation de l'expression (8) pour tous les groupes de renouvellement.

Étant donné Ω, ou bien $\Omega_1, \ldots, \Omega_{k+c-1}$, il est possible de déterminer des expressions pour les estimateurs du maximum de vraisemblance $\hat{\mu}$ et $\hat{\sigma}^2$, qui servent à estimer μ et σ^2 respectivement. De même, il est possible de connaître la matrice des variances-covariances estimée de $\hat{\mu}$, $V(\hat{\mu})$. C'est ce que nous illustrons pour un processus autorégressif du premier degré dans la sous-section 3.2. La fonction de vraisemblance marginale pour $\Omega_1, \ldots, \Omega_{k+c-1}$ est alors définie par l'équation (4), où

$$|\Omega|^{1/2} = \prod_{r=1}^{k+c-1} \Omega_r,$$

$$|X^T \Omega^{-1} X|^{1/2} = V(\hat{\mu})/s^k,$$

$$s^2 = \sum_{r=1}^{k+c-1} \{(n_r \hat{x}_r^T \Omega_r^{-1} \hat{x}_r + (n_r - 1)\,\mathrm{tr}(\Omega_r^{-1} S_r)\}, \tag{9}$$

et $p = k$; dans l'équation ci-dessus, $\hat{x}_r$ est x_r à la différence près que μ est remplacée par l'estimateur du maximum de vraisemblance correspondant.

3.2 Processus autorégressifs du premier degré

Considérons un modèle autorégressif où les unités sont indépendantes les unes des autres mais où il y a corrélation dans le temps pour la même unité. En particulier, supposons le modèle autorégressif du premier degré

$$y_{tj} = \mu_t + \phi\,(y_{t-1,j} - \mu_{t-1}) + \varepsilon_{tj}, \tag{10}$$

où $\varepsilon_{tj} \sim N(0,\sigma^2)$ pour $t = 1, \ldots, k$ and $j = 1, \ldots, N$, et où les ε sont mutuellement indépendants. Le modèle (9), qui correspond essentiellement au modèle de Patterson (1950), est un cas particulier de (7). Comme logarithmique est désignée par $l(\beta,\lambda,\Omega)$ et peut être tirée de l'équation (2). Si les éléments de Ω sont des fonctions d'un paramètre Φ, les paramètres de nuisance λ et β sont l'un et l'autre orthogonaux à Ω, c'est-à-dire

$$-\frac{1}{m}E[\frac{\partial^2 l(\beta,\lambda,\Omega)}{\partial\Phi\partial\lambda}] = 0$$

et

$$-\frac{1}{m}E[\frac{\partial^2 l(\beta,\lambda,\Omega)}{\partial\Phi\partial\beta}] = 0,$$

lorsque chaque élément de Ω est une fonction continue et différentiable de Φ. En outre dans ce cas, la fonction conditionnelle approximative pour Ω, $L_C(\Omega)$ est identique à la fonction marginale, $L_M(\Omega)$, définie en (4) ou en (5). Voir Bellhouse (1990) pour plus de détails.

Il est possible d'évaluer la FM et la FCA définies en (4) ou en (5) pour n'importe quelle matrice Ω en se servant de modèles d'espace d'états à la manière de Harvey et Phillips (1979). Une fois que les calculs récursifs visant à estimer β et σ^2 sont terminés, on peut calculer, pour n'importe quelle matrice Ω donnée, la valeur de S^2 et de $|\Omega|^{1/2}$ au moyen des formules proposées par Harvey et Phillips (1979, équations 5.6 et 6.6, et 4.3 respectivement). On n'a alors qu'à calculer $X^T\Omega^{-1}X$ et son déterminant. La dernière étape du processus récursif de Harvey et Phillips (1979, équation 3.4) permet de déterminer la valeur de $X^T\Omega^{-1}X$.

Supposons, pour ce qui a trait au modèle (1), que β est un vecteur aléatoire défini par l'équation $\beta = W\delta + u$, où W est une matrice $p \times q$ de valeurs connues, δ est un vecteur de paramètres de dimensions $q \times 1$ et $u \sim N(0, \gamma^2\Gamma)$, indépendant de ε. Selon le modèle composé $y = XW\delta + Xu + \varepsilon$, la fonction de vraisemblance logarithmique pour $\delta,\Omega,\Gamma,\gamma^2$, et $\kappa = \sigma^2/\gamma^2$, désignée par $l(\delta,\kappa,\gamma^2,\Gamma,\Omega)$, est définie par l'équation (2), où Ω est remplacé par l'expression $\kappa\Omega + X\Gamma X^T$ et $X\beta$ replaced by $XW\delta$. De la même façon, la fonction de vraisemblance marginale, désignée par $L_M(\kappa,\Gamma,\Omega)$, est définie par les équations (4) et (3), où X est remplacé par XW et Ω, par $\kappa\Omega + X\Gamma X^T$. Ainsi,

$$l_M(\kappa,\Gamma,\Omega) = \{|\kappa\Omega + X\Gamma X^T|^{1/2}|(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW|^{1/2} g^{m-q}\}^{-1}, \qquad (6)$$

où

$$g = y^T(\kappa\Omega + X\Gamma X^T)^{-1}y$$
$$- y^T(\kappa\Omega + X\Gamma X^T)^{-1}XW((XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW)^{-1}(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}y.$$

Or, la dimension d'Ω peut être grande par rapport à celle de Γ; c'est une situation qu'on observe parfois dans l'échantillonnage répété. Comme alternative, on pourrait prendre la fonction de vraisemblance définie en (2), la multiplier par la distribution de β, puis intégrer le tout par rapport à β pour obtenir la fonction de vraisemblance pour les paramètres du modèle des coefficients aléatoires. On aurait ainsi des matrices de la même dimension que Γ.

3. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE RÉPÉTÉ

3.1 Échantillonnage avec renouvellement

Considérons une population finie de N unités qui a été sondée k fois selon un plan avec renouvellement à un niveau. Désignons par y_{tj} la valeur observée pour l'unité de population j au t-ième passage de l'enquête, $j=1, \ldots, N$ and $t=1, \ldots, k$. Au départ, nous supposons que les unités de la population sont indépendantes les unes des autres mais qu'il y a corrélation dans le temps pour la même unité. En particulier, nous supposons que pour n'importe quelle unité j,

$$(y_{1j}, y_{2j}, \ldots, y_{kj})^T \sim N(\mu,\sigma^2\Omega_k), \qquad (7)$$

où Ω_k est une matrice de corrélation $k \times k$ et μ est le vecteur $1 \times k$ de moyennes fixes $(\mu_1, \mu_2, \ldots, \mu_k)^T$.

où le vecteur d'erreurs $\underset{\sim}{\varepsilon} \sim N(0,\sigma^2\Omega)$, où Ω étant la matrice de corrélation, la fonction de vraisemblance logarithmique pour β, σ^2 et Ω est définie comme

$$l(\beta,\sigma^2,\Omega) = -\{m \ln\sigma + (\ln|\Omega|)/2 + (y-X\beta)^T\Omega^{-1}(y-X\beta)/2\sigma^2)\} \qquad (2)$$

Le vecteur des observations **y** est de dimensions $m \times 1$ et le vecteur des coefficients de régression β est $p \times 1$ de sorte que **X** est de dimensions $m \times p$. Pour une valeur donnée de Ω,

$$\hat{\beta} = (X^T\Omega^{-1}X)X^T\Omega^{-1}y$$

et

$$s^2 = y^T\Omega^{-1}y - y^T\Omega^{-1}X\ (X^T\Omega^{-1}X)^{-1}\ X^T\Omega^{-1}y \qquad (3)$$

sont des estimateurs exhaustifs conjoints de β et de σ^2.

On obtient une fonction de vraisemblance marginale pour Ω par une réduction des données **y** aux statistiques exhaustives $\hat{\beta}$ et s^2 et à la statistique ancillaire

$$a = \Omega^{-1/2}(y - X\ (X^T\Omega^{-1}X)^{-1}\ X^T\Omega^{-1}y)/s,$$

où $\Omega^{-1/2}$ est la matrice de dimensions $m \times m$ telle que $\Omega^{-1} = \Omega^{-1/2}\Omega^{-1/2}$. La fonction de vraisemblance marginale de Ω correspond à la distribution marginale de la statistique ancillaire **a** multipliée par le produit des différentielles da_i, $i=1, \ldots, m$. Voir Kalbfleisch et Sprott (1970, équations 6 et 10) pour une analyse globale et une expression générale pour Πda_i. Bellhouse (1978), suivi quelques années plus tard de Tunnicliffe Wilson (1989), a montré que la fonction de vraisemblance marginale pour Ω suivant un modèle normal était définie comme

$$L_M(\Omega) = [\ |\Omega|^{1/2}\ |X^T\Omega^{-1}X|^{1/2}s^{m-p}\ ]^{-1}. \qquad (4)$$

Notons que l'équation (3) est proportionnelle à l'estimateur du maximum de vraisemblance de σ^2 et que Ω et que $s^2(X^T\Omega^{-1}X)^{-1}$ est proportionnelle à la matrice des variances-covariances estimée de l'estimateur du maximum de vraisemblance de β étant donné Ω. Alors, l'équation (4) peut être réécrite

$$L_M(\Omega) = \frac{|\text{est var}(\hat{\beta})|^{1/2}}{s^m|\Omega|^{1/2}}. \qquad (5)$$

Pour définir une fonction de vraisemblance conditionnelle approximative (FCA), il faut tout d'abord transformer les paramètres de manière à obtenir une relation d'orthogonalité entre les paramètres d'intérêt et les paramètres de nuisance, qui peuvent dépendre des premiers. Il y a orthogonalité entre des ensembles de paramètres lorsque la matrice d'information correspondante est une matrice diagonale par blocs, chaque bloc servant lui-même de matrice d'information pour un ensemble de paramètres. La fonction de vraisemblance conditionnelle est liée à la distribution des données **y**, qui dépend elle-même de l'estimateur du maximum de vraisemblance des paramètres de nuisance pour des valeurs déterminées des paramètres d'intérêt. On obtient la fonction de vraisemblance conditionnelle approximative en appliquant deux approximations à cette distribution conditionnelle. Voir Cox et Reid (1987, section 4.1) pour une analyse des calculs. Par exemple, posons θ comme le vecteur des paramètres d'intérêt et Λ, qui dépend possiblement de θ, comme le vecteur des paramètres dérangeants orthogonal à θ. La fonction de vraisemblance complète pour les paramètres θ et Λ est désignée par $L(\theta,\Lambda)$ tandis que la fonction-profil pour θ, $L(\theta,\hat{\Lambda})$ équivaut à la fonction de vraisemblance ordinaire à la différence près que Λ est remplacé par l'estimateur du maximum de vraisemblance correspondant. La fonction de vraisemblance conditionnelle approximative pour θ est

$$L(\theta,\hat{\Lambda})\ |\ I(\theta,\hat{\Lambda})\ |^{1/2},$$

où $I(\theta,\hat{\Lambda})$ est la matrice d'information observée pour Λ à une valeur déterminée de θ. Voir Cox et Reid (1987, équation 10).

À la suite de Cruddas et coll. (1989), Bellhouse (1990) a proposé pour le modèle (1) la transformation $\lambda = \ln\sigma + (\ln|\Omega|)/(2m)$, β demeurant inchangé. Suivant ces nouvelles conditions, la fonction de vraisemblance

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

FONCTIONS DE VRAISEMBLANCE MARGINALES ET FONCTIONS DE VRAISEMBLANCE CONDITIONNELLES APPROXIMATIVES POUR L'ÉCHANTILLONNAGE RÉPÉTÉ

D.R. Bellhouse[1]

RÉSUMÉ

L'auteur définit des fonctions de vraisemblance marginales et des fonctions de vraisemblance conditionnelles approximatives pour les paramètres de corrélation d'un modèle de régression linéaire normal à erreurs corrélées; il suppose tantôt des paramètres de régression fixes, tantôt des coefficients aléatoires. Ces fonctions de vraisemblance peuvent être évaluées à l'aide de modèles d'espace d'états. L'auteur se sert du principe de vraisemblance pour déterminer des fonctions marginales et conditionnelles pour les paramètres de corrélation dans un plan d'échantillonnage répété (échantillonnage aléatoire simple et plans plus complexes).

MOTS CLÉS: Inférence fondée sur la vraisemblance; échantillonnage dans le temps; modèles ARMA; modèles d'espace d'états.

1. INTRODUCTION

On a proposé pour la première fois les fonctions de vraisemblance marginales (FM) comme une méthode générale pour éliminer les paramètres de nuisance de la fonction de vraisemblance (Fraser, 1967; Kalbfleisch et Sprott, 1970). Les fonctions de vraisemblance conditionnelles approximatives (FCA) ont été définies dans le même but par Cox et Reid (1987). Ceux-ci affirment que la FCA est préférable à la fonction-profil de vraisemblance, que l'on obtient en remplaçant les paramètres de nuisance dans la fonction de vraisemblance par l'estimation la plus vraisemblable correspondante lorsque les paramètres d'intérêt sont connus. Bellhouse (1990) a démontré l'équivalence de la FM et de la FCA pour des paramètres de corrélation suivant un modèle normal. S'inspirant de l'étude de Cox et Reid, Cruddas et coll. (1989) ont établi une FCA pour les paramètres de corrélation dans plusieurs petites séries de processus autorégressifs du premier degré ayant la même variance et les mêmes paramètres d'autocorrélation. Ils ont montré par une étude de simulation que l'estimateur fondé sur la FCA était beaucoup moins biaisé que l'estimateur du maximum de vraisemblance fondé sur la fonction profil et qu'il correspondait à un intervalle de confiance plus étendu.

La question que traitent Cruddas et coll. (1989) est illustrée dans les enquêtes répétées. Afin de réduire le fardeau de réponse des personnes qui participent à ce genre d'enquêtes, on fait en sorte qu'elles ne fassent pas trop longtemps partie de l'échantillon. À chaque tenue de l'enquête, l'échantillon est formé de personnes qui en sont au moins à leur seconde participation et d'autres qui en sont à leur première participation. Les données recueillies à cette occasion sur chaque personne sont normalement modélisées à l'aide d'un processus autorégressif de moyennes mobiles (ARMA); voir Binder et Hidiroglou (1988) pour une étude de l'application des modèles de séries chronologiques pour l'échantillonnage répété. En outre, à cause du fardeau de réponse, la série des observations pour chaque individu est courte. Si l'on suppose que les moyennes de modèle sont différentes d'une fois à l'autre, la dimension de l'espace des paramètres augmentera avec le temps de sorte que l'estimateur du maximum de vraisemblance des paramètres pourrait être biaisé et non consistant. C'est pourquoi il est utile de définir des FM et des FCA suivant des modèles ARMA.

Dans la section 2, nous déterminons les FM et FCA pour les paramètres de corrélation suivant un modèle normal. Nous appliquons ensuite les résultats de cette section aux enquêtes à passages répétés avec plan d'échantillonnage aléatoire simple. Enfin dans la section 4, nous présentons plusieurs méthodes qui permettent d'appliquer ces fonctions de vraisemblance à des plans de sondage complexes.

2. FONCTION DE VRAISEMBLANCE MARGINALE ET FONCTION DE VRAISEMBLANCE CONDITIONNELLE APPROXIMATIVE POUR DES PARAMÈTRES DE CORRÉLATION SUIVANT UN MODÈLE NORMAL

Pour le modèle linéaire

$$y = X\beta + \varepsilon \tag{1}$$

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London (Ontario), Canada N6A 5B9

Une des alternatives à la technique des données manquantes est d'étudier les commerces sans lecteurs optiques à l'aide d'une autre méthode de collecte des données. Selon les données à recueillir, cela peut comprendre une vérification des dossiers de l'entreprise ou une entrevue avec le personnel, par téléphone, par courrier ou en personne. Cette méthode est plus précise que la méthode d'imputation, mais elle est plus coûteuse et plus longue et les problèmes de gestion créés par deux méthodes de collecte des données sont importants.

En dernier lieu, nous traiterons de l'abandon de la lecture optique par les commerces. Ce genre de cas est assez rare et il n'est traité que pour s'assurer de bien couvrir tout le sujet.

Posons $U_{ijb} \varepsilon s_A$, c'est-à-dire que i est un commerce avec un lecteur optique dans l'échantillon. Notons que U_{ijb} peut être un commerce qui possède un lecteur optique depuis le début de l'étude ou un commerce qui a adopté la lecture optique après avoir fait partie de l'échantillon comme commerce sans lecteur optique conformément à la règle 5.

> **Règle 6. Au moment où U_{ijb} abandonne la lecture optique, il doit être soustrait de s_A, ajouté à s_B et traité par les méthodes de données manquantes, comme dans la règle 5. Les formules standard doivent être appliquées pour compléter la série de données. Pour simplifier le procédé et le travail d'échantillonnage, la méthode choisie doit être identique à celle choisie pour traiter les commerces qui adoptent la lecture optique.**

Dans le cas inhabituel où un commerce utilise la lecture optique de façon intermittente, il faut le traiter en appliquant les règles 5 ou 6 selon le cas, en mettant chaque fois à jour les échantillons s_A et s_B.

BIBLIOGRAPHIE

Ernst, L. (1989) "Weighting Issues for Longitudinal Household and Family Estimates", dans *Panel Surveys*, édité par Kaspryzk, D., Duncan, G., Kalton, G., et Singh, M.P., Wiley, NY.

Hanson, R.H. (1978) *"The Current Population Survey: Design and Methodology"* Technical Paper 40, United States Bureau of the Census, Washington, DC.

Laurini, R. (1987), "Manipulation of Spatial Objects by a Peano Tuple Algebra," University of Maryland Technical Report CS-TR-1893, College Park, MD.

Peano, G. (1908) "La Curva di Peano nel 'Formulario Mathematico." Dans "Opere Scelte di G. Peano," p. 115-116, Vol. I. Edizioni Cremonesi, Roma, 1957.

Progressive Grocer (1989) "56th Annual Report of the Grocery Industry 1989," Vol. 68, no. 4, part 2, Stamford CT.

Rao, J.N.K. et Graham, J.R. (1964) "Rotation Designs for Sampling on Repeated Occasions", *Journal of the American Statistical Association*, 59, 492-509.

Saalfeld, A.; Fifield, S.; Broome, F.; et Meixler, D. (1988) "Area Sampling Strategies and Payoffs using Modern Geographic Information System Technology," non publié, United States Bureau of the Census, Washington, DC.

Sirken, M. (1970) "Household Surveys with Multiplicity", *Journal of the American Statistical Association*, 65, 257-266.

Wolter, K.M. (1979) "Composite Estimation in Finite Population", *Journal of the American Statistical Association*, 74, 604-613.

Wolter, K.M. (1986) *Introduction to Variance Estimation*, Springer Verlag, NY.

Wolter, K.M. et al (1976) "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys, "*Proceedings of the Business and Economic Statistics Section*, American Statistical Association, Alexandria, VA.

La solution a) est la plus simple et la plus claire. À part le problème des ouvertures, elle n'est pas biaisée et permet un bon calcul de l'estimateur de variance. En raison du problème des ouvertures, toutefois, cette solution peut avoir un effet négatif sur la capacité de l'échantillon à mesurer les tendances. Si les fermetures se produisent pendant la première semaine d'un cycle, on peut remarquer le léger déclin dans le temps, non pas en raison d'un changement fondamental des conditions économiques, mais simplement parce que l'échantillon tient compte des fermetures et non des ouvertures. La solution b) est une solution à court terme pour la mesure adéquate des tendances. La notion essentielle est qu'en donnant une valeur aux commerces fermés, nous compensons de façon implicite pour toutes les ouvertures qui ont pu se produire depuis le dernier cycle de mise à jour. Cette solution n'est pas particulièrement élégante et il est difficile de la justifier. Cependant, l'histoire nous montre que l'univers des commerces est stable à court terme. Les fermetures sont souvent associées à des ouvertures ou compensées par celles-ci et la taille nette de la population reste à peu près constante à court terme. Le United States Bureau of the Census a utilisé cette solution dans son enquête sur le commerce de gros avec des cycles de mise à jour trimestriels et une collecte des données mensuelle. Voir Wolter et Coll. (1976).

4.3 Lecture optique

Dans cette dernière section, nous présentons les règles de mise à jour des échantillons dans le cas des commerces qui adoptent le système de lecture optique ou qui l'abandonnent. Bien entendu, ce problème ne se pose pas dans les sondages où la collecte des données se fait d'une façon différente.

Nous parlerons en premier lieu des commerces qui adoptent le système de lecture optique. Il y a deux cas à considérer: (i) le système utilisé par tous les commerces est connu avant l'échantillonnage et (ii) le système utilisé n'est connu qu'après l'échantillonnage et uniquement pour les commerces sélectionnés.

Le cas i) est relativement simple. Voici la règle à suivre:

> **Règle 4. Ne pas inclure les commerces n'utilisant pas la lecture optique dans l'échantillon. Ne choisir l'échantillon que parmi les commerces qui utilisent la lecture optique. Si un commerce adopte le lecteur optique, le traiter comme une ouverture et le soumettre à l'échantillonnage des ouvertures. Avant leur conversion, les commerces n'ayant pas de lecteur optique doivent être traités par imputation ou par une autre technique pour données manquantes.**

Selon cette règle et les données connues (ex: lecture optique ou non), tous les fonds alloués pour le sondage doivent être consacrés à l'échantillonnage des commerces avec lecteurs optiques. Aucun montant ne doit être versé pour le sondage des commerces sans lecteurs optiques. Malheureusement, cette règle ne s'applique pas au cas ii).

Pour résoudre le cas ii), il faut introduire une notation supplémentaire. Posons A comme étant l'ensemble des commerces avec lecteurs optiques et B l'ensemble des commerces sans lecteurs optiques; A B représente toute la population. Posons s pour l'échantillon de commerces sélectionnés et soient $s_A = s \quad A$ et $s_B = s \quad B$.

Par hypothèse, s_A et s_B ne sont connus qu'après l'étape initiale de travail sur le terrain. Chacun de ces ensembles varie avec le temps mais toute référence au temps a été supprimée, afin de simplifier la notation.

L'échantillon S_A doit être mis à jour conformément aux règles indiquées dans cet article pour les ouvertures et les fermetures. Dans le cas de s_B, il faut établir de nouvelles règles. Voici une règle qui traite les commerces de s_B comme non répondants:

> **Règle 5. Au temps t, imputer au commerce $U_{ijb} \epsilon s_B$ la valeur $\hat{y}_{tijb} = x_{tijb} y_{At} / x_{At}$, où x_{tijb} est la valeur d'une variable auxiliaire pour le commerce U_{ijb}, y_{At} est le total de la variable d'estimation pour l'échantillon s_A et x_{At} est le total correspondant pour la variable auxilialre. De plus, l'imputation peut se faire par substitution, par appariement "Hot Deck" ou par un autre moyen. Maintenant, en supposant que la série de données est complète, appliquer l'estimateur standard du paramètre d'intérêt. Au moment où U_{ijb} adopte la lecture optique, il doit encore être soustrait de s_B et ajouté à s_A et l'estimation doit être effectuée à l'aide de l'estimateur standard appliqué à la série de données complétée.**

Selon la règle 5, la taille efficace de l'échantillon est réduite en raison de la variance d'imputation associée avec $\hat{y}_{tijb}$. La substitution permet de conserver une taille d'échantillon supérieure aux autres règles, mais c'est la méthode la plus coûteuse. Toutes les règles nécessitent un travail restreint mais continuel sur le terrain pour vérifier si $U_{ijb} \epsilon s_B$ a adopté ou non la lecture optique.

Pour un cycle de mise à jour à un temps donné t', les règles 1 ou 1A peuvent être utilisées pour la mise à jour de l'échantillon. Les nouveaux commerces sont automatiquement placés dans le segment approprié à l'aide de leurs valeurs de Peano et l'indice b reflète cet ordre pour chaque cycle. Pour bien représenter ces notions, il aurait fallu ajouter la notion de temps en indice aux valeurs de U, de B, de p et de Π, mais nous ne l'avons pas fait pour faciliter la notation. Les formules des estimateurs de totaux $\hat{Y}_{t'i}$ et $\hat{Y}_{Rt'i}$ et de leur variance restent valides pour chaque valeur de t'.

4.2 Mise à jour pour la fermeture de commerces

Les règles de mise à jour d'un échantillon dans le temps doivent suivre un important principe général. Elles doivent traiter de la même façon les unités sélectionnées et les unités non sélectionnées. Dans le cas de fermeture de commerces, ce principe signifie que toutes les fermetures, celles qui se produisent à l'intérieur comme à l'extérieur de l'échantillon doivent être traitées de la même façon dans tous les procédés de mise à jour de l'échantillon. Si ce principe n'est pas respecté, les estimateurs seront biaisés et ce biais risque de s'accumuler avec le temps.

Dans les paragraphes suivants, nous décrivons des procédés de mise à jour de l'échantillon en cas de fermeture de commerces qui suivent ce principe essentiel. Nous verrons deux cas: (1) les fermeture ne sont pas connues pour toute la population et (2) les fermetures sont connues pour toute la population.

Pour le cas i), nous suggérons la règle 2

> **Règle 2. Toutes les fermetures de commerces dans l'échantillon sont connues. Les commerces fermés doivent demeurer dans l'échantillon mais leur valeur doit être de 0 (c.-à-d. y = 0) au moment de la mise à jour.**

Cette règle permet de faire une estimation non biaisée des totaux de population de l'univers de l'enquête. Les fermetures peuvent faire augmenter la variance de l'estimateur, et les estimateurs de variance vont refléter cette augmentation si les commerces fermés sont conservés dans l'échantillon avec une valeur de 0.

Pour le cas ii), nous suggérons la règle 3.

> **Règle 3. Retirer tous les commerces fermés de la population au moment de la mise à jour suivante. Seuls les commerces ouverts seront soumis à l'échantillonnage, y compris les nouveaux commerces.**

La règle 3 fait varier le nombre de commerces (B_{ij}) dans les segments où il y a eu des fermetures, à moins que le nombre de nouveaux commerces soit égal au nombre de fermetures. En fait, les valeurs de B et de p vont obligatoirement changer dans les segments où il y a eu des fermetures et aucune ouverture de commerce. Par conséquent, un commerce de remplacement sera choisi à l'intérieur d'un segment chaque fois que le commerce échantillonné de ce segment fermera ses portes. Un commerce de remplacement peut être choisi même si le commerces échantillonné est encore ouvert.

Deux autres problèmes doivent être étudiés dans le cas d'une fermeture de commerce. Le premier concerne la coordination des mises à jour des ouvertures et des fermetures. L'ouverture et la fermeture des commerces se produit naturellement à des intervalles irréguliers, selon la situation économique et la croissance de la population. Pendant certaines périodes, il peut n'y avoir ni ouverture ni fermeture de commerces. Pendant d'autres périodes, certains commerces peuvent faire leur apparition sans qu'il n'y ait de fermeture ou vice versa. Tandis que pendant d'autres périodes, il y aura à la fois des ouvertures et des fermetures. En théorie, il est possible d'utiliser différents cycles de mise à jour pour les ouvertures et les fermetures de commerces. Par exemple, la mise à jour peut être bimestrielle, mais alterner pour l'ouverture et la fermeture. Cette approche a l'avantage de régulariser la charge de travail. Cependant, les cycles alternatifs nuisent à la capacité de l'échantillon de bien mesurer les tendances, créant un effet en dent de scie dans la série chronologique des commerces puisque les ouvertures sont incluses dans l'échantillon avant que les fermetures en soient éliminées. Nous recommandons de faire les mises à jour au même moment afin de préserver les tendances.

Le deuxième problème est celui de la période qui s'écoule entre la fermeture du commerce et la mise à jour suivante. Ce problème ne se pose que lorsque la fréquence des mises à jour est inférieure à celle de la collecte des données. Si les deux sont faites en même temps, il n'y a pas de problème. Si la mise à jour est moins fréquente que la collecte, il y a deux solutions:

a) Éliminer les commerces fermés de l'échantillon au moment où la fermeture est connue (pour être plus précis du point de vue statistique, cela signifie que les disparitions sont conservées dans l'échantillon avec une valeur de zéro.)

b) Garder les commerces fermés dans l'échantillon en leur attribuant une valeur jusqu'au moment de la mise à jour suivante.

Règle 1A. Un nouveau commerce pourra faire partie d'un sous-échantillon si et seulement si sa valeur de Peano fait partie d'un segment de Peano sélectionné, On associe les commerces U_{ij1}, U_{ij2}, ... U_{ijBij} aux probabilités P_{ij1}, P_{ij2}, ..., P_{ijBij} où $p_{ijb} > 0$ et $\Sigma p_{ijb} = 1$. On choisit ensuite un des commerces selon cette mesure de probabilité. Le sous-échantillonnage est indépendant d'un segment choisi à l'autre. Les nouveaux commerces dont les valeurs de Peano font partie d'un segment non sélectionné ne sont pas conservés.

Les probabilités dans la règle 1A peuvent être égales ou inégales. Si elles sont inégales, elles peuvent être définies par rapport à une mesure préliminaire de taille ou dans le but d'accélérer ou de retarder le remplacement d'un échantillon.

Les principaux objectifs de mise à jour sont remplis par la règle 1A. La règle permet de maintenir l'équilibre géographique dans le temps puisqu'il n'y a qu'une unité choisie à partir de chaque segment choisi à l'origine, segments qui sont eux-mêmes équilibrés géographiquement en raison du plan d'échantillonnage systématique. Deuxièmement, la règle maintient constante la taille de l'échantillon dans le temps puisqu'il n'y a toujours qu'un seul commerce choisi dans chacun des segments d'origine. Troisièmement, la règle est en accord avec les principes de l'échantillonnage probabiliste où les probabilités d'inclusion doivent être connues et différentes de zéro; ainsi des estimateurs non biaisés de la population totale peuvent être obtenus. Enfin, par un choix approprié de P_{ijb}, il est possible de contrôler la distorsion dans les tendances annuelles.

Les probabilités de sélection inconditionnelles sont calculées ainsi:

$$\pi_{ijb} = k^{-1} p_{ijb}$$

pour $b = 1, \ldots, B_{ij}$. Cela signifie que π_{ijb} est égal à la probabilité de choisir l'unité primaire d'échantillonnage multipliée par la probabilité conditionnelle de choix du commerce, étant donné l'U.P.É. sélectionnée.

Posons maintenant $Y_{t'ijb}$ comme valeur de l'unité U_{ijb} et $Y_{t'ij+}$ pour le total de la $(i,j)^{\text{ème}}$ U.P.É.. L'estimateur non biaisé de la population totale Y_t se calcule alors comme suit:

$$\hat{Y}_{t'i} = \sum_{j=1}^{n_i} y_{t'ijb}/\pi_{ijb}$$

où $y_{t'ijb}$ est la valeur de l'unité choisie à partir du $(i,j)^{\text{ème}}$ segment sélectionné, avec une variance de:

$$\text{Var}\ \{\hat{Y}_{t'1}\} = \frac{1}{k} \sum_{i=1}^{k} (k \sum_{j=1}^{n_i} Y_{t'ij+} - Y_{t'})^2 + K \sum_{i=1}^{k} \sum_{j=1}^{n_i} \sigma^2_{t'ij}, \tag{1}$$

où

$$\sigma^2_{t'ij} = \sum_{b=1}^{B_{ij}} P_{ijb} \left(\frac{y_{t'ijb}}{P_{ijb}} - Y_{t'ij+}\right)^2.$$

Le premier terme à la droite de l'équation (1) est la variance due à l'échantillonnage des segments. C'est la variance initiale puisque c'est la variance qui s'appliquait au moment du choix de l'échantillon original. Le second terme à la droite est la variance due au sous-échantillonnage à l'intérieur des segments. Noter que la valeur $\sigma^2_{t'ij}$ disparaît pour tout segment où il n'y a pas eu de sous-échantillonnage de nouveaux commerces. Noter aussi que la variance du sous-échantillonnage est minimale lorsque, pour tout i et j, les probabilités p_{ijb} sont proportionnelles à $Y_{t'ijb}$. Dans ce cas, la composante de la variance à l'intérieur des segments s'annule. Pour une application réelle, toutefois, cette condition de proportionnalité n'est satisfaite qu'en partie.

Comme d'habitude, une approximation de Taylor du premier ordre peut être utilisée pour découvrir la variance de l'estimateur par le quotient. Voir Wolter, (1986) pour les techniques appropriées d'estimation de la variance de l'estimateur non biaisé $\hat{Y}_{t'i}$ et de l'estimateur par le quotient $\hat{Y}_{Rt'i}$.

À mesure que le temps passe, il faut mettre à jour l'échantillon pour refléter les ouvertures de commerces et les autres changements de la population. Il peut être bon de prévoir une mise à jour à des intervalles réguliers afin de faciliter le travail. Ces intervalles sont appelés cycles de mise à jour. De tels cycles peuvent être mensuels, bimestriels, trimestriels ou selon les exigences d'une application en particulier. Les facteurs à tenir en ligne de compte pour l'établissement du cycle de mise à jour sont le coût de la mise à jour, la précision des estimateurs de niveau et de tendance recherchée et les besoins des clients ou des utilisateurs des données.

De façon générale, un échantillonnage mis à jour fréquemment sera plus coûteux, plus précis et plus apprécié des clients qu'un échantillon mis à jour moins souvent.

Supposons qu'un échantillon original de commerces est choisi avec probabilités égales systématiquement dans la liste de Peano au temps $t = 0$. Soit U_{ij}, le commerce j dans l'échantillon systématique i, pour $i = 1, \ldots, k$ et $j = 1, \ldots, n_i$; k étant l'intervalle d'échantillonnage et n_i la taille de l'échantillon aléatoire i. Si $N = nk + r$, $r < k$, r échantillons seront de taille $n_i = n + 1$ et $k - r$ échantillons seront de taille n. Dans les paragraphes suivants, l'indice i est utilisé pour représenter l'échantillon choisi.

Soit P_{ij} la valeur de Peano associée à U_{ij}. Prenons P_L et P_u comme, respectivement, les plus petites et plus grandes valeurs de Peano possibles pour le marché à l'étude. Ainsi,

$$P_L \leq P_{11} < P_{21} < \ldots < P_{k1} < P_{12} < \ldots < P_{ij} < \ldots < P_{kn_k} \leq P_U.$$

Nous supposons que chaque commerce n'a qu'une seule situation géographique et donc une seule valeur de Peano.

Soit Y_{tij} la valeur d'une certaine caractéristique de U_{ij} au temps t. Un estimateur standard, non biaisé de la population totale, Y_t, est le suivant:

$$\hat{Y}_{ti} = k \sum_{j=1}^{n_i} y_{tij},$$

Tandis que l'estimateur par le quotient est donné par:

$$\hat{Y}_{Rti} = \hat{Y}_{ti}\, X_t \,/\, \hat{X}_{ti},$$

où la variable X est une mesure de la taille et X_t et X_{ti} sont analogues respectivement, à Y_t et à $\hat{Y}_{ti}$.

Définissons maintenant N segments de Peano, S_{ij}, en divisant l'échelle $[P_L, P_U]$ aux N valeurs de P_{ij}. Posons $S_{ij} = [P_{ij}, P_{i+1,\, j}]$ où $P_{k+1,\, j}$ représente $P_{1,\, j+1}$. Une définition spéciale est nécessaire pour le dernier segment. $S = [P_{kn_k}, P_U] \cup [P_L, P_{11}]$ afin que toute l'échelle de Peano $[P_L, P_U]$ soit couverte par les N segments. Cette définition spéciale, qui permet de définir l'échelle de Peano comme si elle était un cercle, est nécessaire pour s'assurer que la probabilité de sélection de tous les nouveaux commerces est différente de zéro. D'autres plans de segmentation peuvent être utilisés sans nuire aux propriétés statistiques du systèmes de mise à jour.

Notre système de mise à jour est basé sur les segments de Peano. Il permet de sélectionner les segments de façon systématique et de faire un sous-échantillonnage des commerces à l'intérieur des segments choisis. Ainsi, c'est le segment qui est l'unité primaire d'échantillonnage (U.P.É.) et non le commerce. Bien entendu, au moment du choix de l'échantillon initial, chaque segment ne compte qu'un seul commerce.

4.1 Échantillonnage des nouveaux commerces

Il est possible qu'un ou plusieurs commerces voient le jour à une période future t' du sondage. On assigne alors à chaque nouveau commerce une valeur de Peano unique qui fait partie d'un segment de Peano. La valeur de Peano permet de placer les nouveaux commerces automatiquement à l'endroit approprié dans la liste ordonnée de l'univers de l'enquête.

La règle la plus simple pour l'échantillonnage des nouveaux commerces est la suivante:

> **Règle 1: Un nouveau commerce est inclus dans l'échantillon si et seulement si sa valeur de Peano fait partie d'un segment de Peano sélectionné. Les nouveaux commerces dont la valeur fait partie d'un segment non sélectionné ne sont pas inclus dans l'échantillon.**

Selon cette règle, la probabilité de sélection d'un nouveau commerce est de $1/k$. Cela se produit parce que la probabilité de sélection du segment, qui est unique, est de $1/k$. Malheureusement, la règle 1 ne permet pas de contrôler la taille de l'échantillon dans le temps.

Pour contrôler la taille de l'échantillon, nous proposons une forme de sous-échantillonnage à l'intérieur de l'unité primaire. Supposons que $U_{ij1}, U_{ij2}, \ldots, U_{ijB_{ij}}$ soient les commerces du segment S_{ij}. Le commerce original est U_{ij1} alors que les commerces $U_{ij2}, U_{ij3}, \ldots, U_{ijB_{ij}}$ sont les nouveaux commerces dans l'ordre de Peano. Le nombre de nouveaux commerces $(B_{ij} - 1)$ dans un segment donné sera dans la plupart des cas 0, 1 ou 2.

Ainsi, nous pouvons effectuer le sous-échantillonnage de la façon décrite ci-après.

FIGURE 2.
Courbe de Peano sur 1024 Points

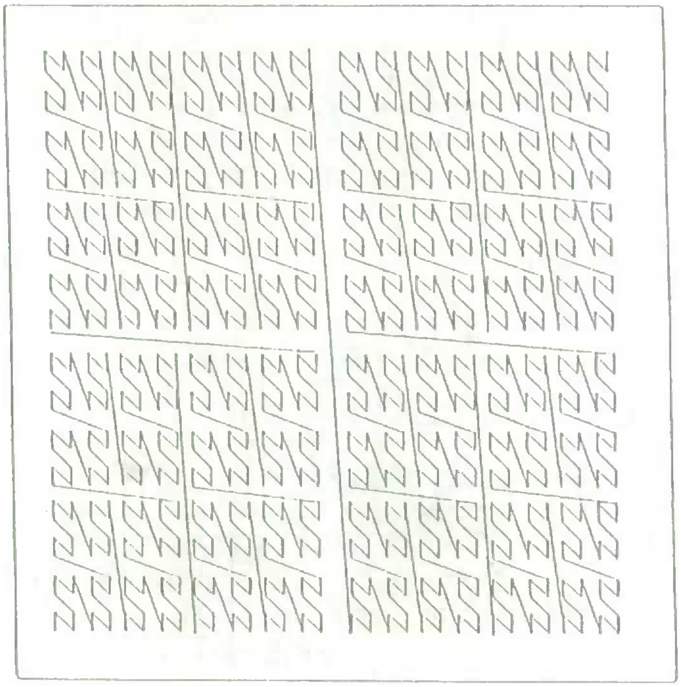
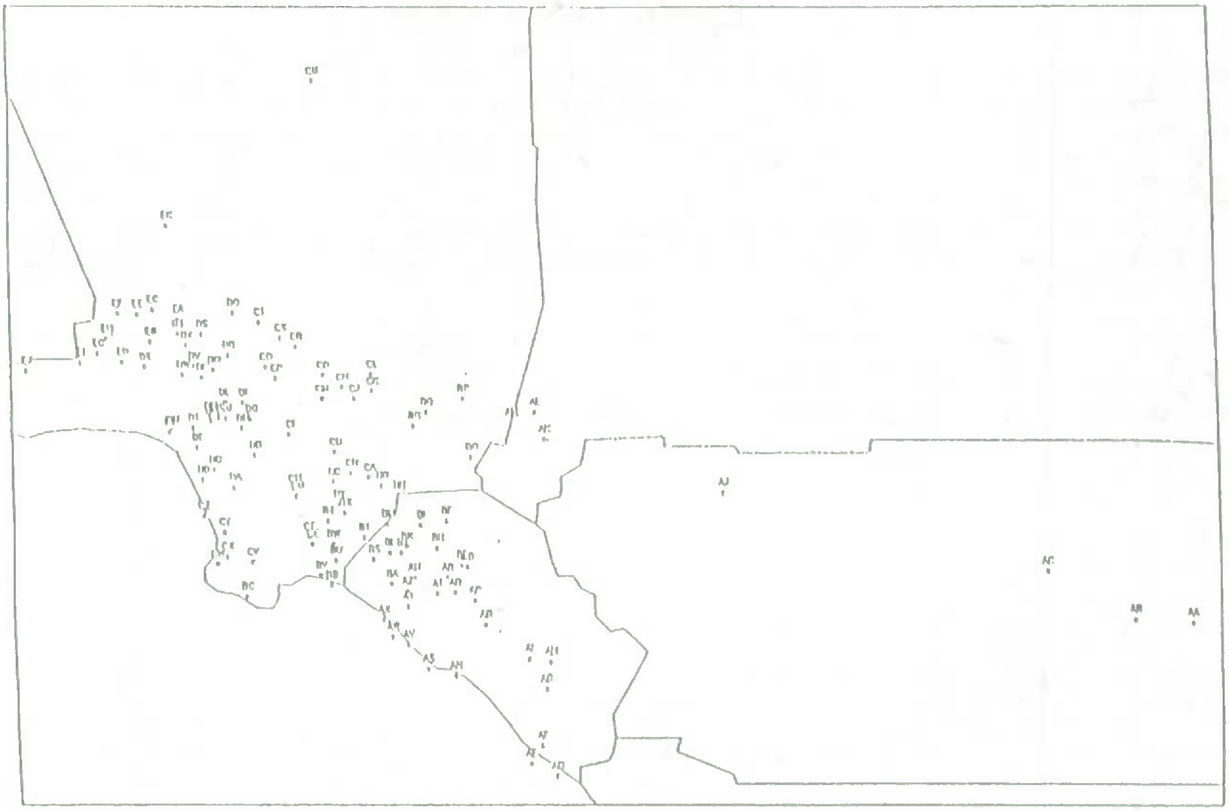
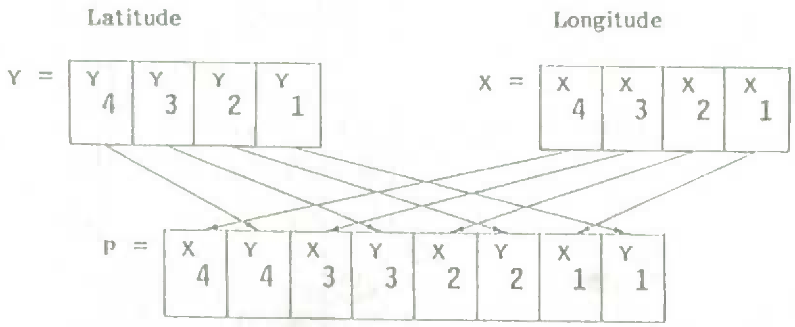


FIGURE 3
Chaîne de commerces de détail dans l'ordre de Peano



$P = X_k\ Y_k\ \ldots\ X_3\ Y_3\ X_2\ Y_2\ X_1\ Y_1$. Voir la figure 1 pour un exemple dans le cas où k = 4. Noter comme il est simple de calculer la valeur de P.

Figure 1. Calcul de la valeur de Peano par entrecroisement de nombres binaires



Étant donné (pour tout k fini) des coordonnées de latitude et de longitude de k chiffres, le "point" spatial représenté par la valeur de P est en fait un carré dans R^2. À mesure que k augmente, la taille des carrés diminue. En fait, lorsque k tend vers l'infini, la valeur de P tend à représenter un point spécifique dans R^2.

La courbe de remplissage créée par les valeurs de Peano prend la forme d'un N récursif. La courbe en N est illustrée à la figure 2 sur une grille de 1024 points. Cette figure montre bien l'aspect récursif des images fractales.

La courbe en N passe une fois et une fois seulement par chaque point de l'espace, les points étant des carrés dont la taille est déterminée par le nombre de chiffres des coordonnées de longitude et de latitude. L'ordre des points sur la courbe (ordre de Peano) préserve à un haut degré la contiguité géographique, ce qui facilite les études sur la proximité. L'ordre de Peano entraîne quelques discontinuités géographiques (saut du point 516 au point 517 dans la figure 2) comme toute correspondance entre R^2 et R^1.

Dans l'application visée, les établissement économiques sont placés sur une liste dans l'ordre de Peano à l'aide de leurs coordonnées de latitude et de longitude. Des échantillons aléatoires d'établissements peuvent être tirés systématiquement de cette liste. Comme les coordonnées géographiques de la terre sont des données stables, la position des nouveaux établissement sur la liste peut être établie sans équivoque. Ces établissements peuvent donc aussi être échantillonnés.

Pour illustrer cette application, voyons la figure 3 qui montre une chaine de commerces de détail aux États-Unis. Chaque établissement est représenté par un code à deux lettres. L'ordre de Peano des établissements suit l'ordre alphabétique des codes.

Le chapitré suivant traite du système de mise à jour basé sur la liste des établissements en ordre de Peano.

Figure 2. Courbe de Peano sur 1024 points.

Figure 3. Chaîne de commerces de détail dans l'ordre de Peano.

4. RÈGLES DE MISE À JOUR D'UN ÉCHANTILLON

Dans les paragraphes suivants, nous décrivons un système de mise à jour des échantillons de commerces de détail en tenant compte des ouvertures, des fermetures, de l'adoption ou de l'abandon de la lecture optique et des autres changements qui peuvent se produire dans l'ensemble des commerces de détail. Comme nous l'avons dit plus tôt, ce système a été développé pour les applications de la A.C. Nielsen Company.

Supposons une strate d'échantillonnage donnée et arbitraire, de taille N où les commerces sont répartis selon l'ordre de Peano. Par exemple, une strate peut comprendre tous les commerces dans un marché métropolitain donné, comme Vancouver ou Montréal. La répartition par les valeurs de Peano est bien adaptée au système de mise à jour ci-après. D'autres types de répartition peuvent être utilisés, à condition qu'ils soient stables dans le temps et qu'ils puissent relier R^2et R^1 de façon à conserver la contiguité géographique et à donner à chaque nouveau commerce une position unique dans la liste ordonnée des commerces.

Les rapports NSUS comprennent le total estimé des ventes de chaque article et groupe d'articles pour chacun des marchés et pour l'ensemble des États-Unis. Un estimateur par le quotient est utilisé et la variable auxiliaire est le volume des ventes. Le volume des ventes est le total des ventes de tous les articles d'un magasin, habituellement sur une période d'un an. Cette valeur est en étroite corrélation avec les ventes des articles individuels. De plus, le rapport NSUS comprend des estimations des ventes et des taux de vente en fonction de la commercialisation des produits ainsi que des estimations des tendances annuelles de vente.

L'échantillon Scantrack doit être continuellement mis à jour, parce que l'univers de l'enquête composé d'environ 30 500 supermarchés n'est pas statique. Lors d'une période récente de 12 mois, environ 2 200 nouveaux supermarchés ont vu le jour et 2 450 ont fermé les portes. De plus, 170 établissements ont été reclassifiés. La reclassification peut être causée par différents facteurs. Certains petits magasins d'alimentation font leur entrée dans l'univers Scantrack lorsque leur volume de ventes devient supérieur à 2 millions de dollars par année, ce qui les fait entrer dans la catégorie des supermarchés. Un magasin peut déménager, chager de nom ou devenir plus grand. Certains supermarchés peuvent changer de statut et devenir des hypermarchés, des magasins-entrepôts ou un autre type de supermarché non traditionnel. En 1979, les 3 800 hypermarchés et magasins-entrepôts étaient responsables de 17% des ventes totales des supermarchés. En 1988, leur nombre est passé à 9000 et ils étaient responsables de près de 50% des ventes totales (Progressive Grocer 1989). Dans certains cas, des supermarchés ou même des chaînes entières de supermarchés sont acquis par une autre organisation, ce qui influe sur la définition des strates.

En plus des changements de population, le manque de données ou leur inexactitude peut entraîner le remplacement de certaines unités d'échantillonnage. Certains supermarchés sélectionnés ne possèdent pas de lecteurs optiques ou ont des lecteurs optiques qui ne sont pas compatibles. Si un supermarché ne peut fournir les données utiles de façon constante, il doit être éliminé de l'échantillon. Dans certains cas, la demande de changement de l'échantillon provient de la chaîne elle-même. Occasionnellement, un détaillant peut simplement refuser de coopérer.

Les principaux objectifs du système de mise à jour pour l'échantillon Scantrack sont les suivants: 1) maintenir l'équilibre géographique de l'échantillon dans le temps, 2) maintenir la taille de l'échantillon dans le temps, 3) s'assurer que les principes de l'échantillonnage probabiliste sont respectés afin d'éviter que les estimateurs des ventes totales ne soient biaisés et 4) s'assurer que les changements dans l'échantillon ne faussent pas excessivement les estimations des tendances annuelles.

L'équilibre géographique est un autre terme pour équilibre socio-économique. Comme différents quartiers ont différentes habitudes de consommation, l'équilibre géographique est important pour que le plan d'échantillonnage soit efficace (i.e.: faible variabilité de l'échantillon) pour un grand nombre d'articles. De plus, les clients considèrent que l'équilibre géographique est un facteur important pour un échantillonnage adéquat.

La diminution de la taille d'un échantillon fait augmenter l'erreur-type des estimateurs tandis que son augmentation accroît le coût du sondage. Les deux situations sont à éviter. De plus, les contrats passés avec les organisations indiquent la taille des échantillons et les paiements et tout changement doit être renégocié. Cette situation est aussi à éviter.

Toutes les applications utilisant les données Scantrack requièrent des estimateurs efficaces et non biaisés des ventes totales. Les fabricants et les commerçants utilisent ces données tous les jours pour prendre des décisions: quelle quantité de produits faut-il fabriquer, quelle quantité faut-il distribuer, quelle quantité garder en stock, comment organiser les étalages, etc.

Les clients ont également besoin d'estimations fiables sur les tendances annuelles pour la gestion de leurs affaires. Ces données permettent aux fabricants d'évaluer la santé de leur entreprise. Il est profitable, à la fois pour le commerçant et pour le fabricant de connaître le rendement à long terme de toutes les marques importantes dans toutes les catégories de produits.

Le système de mise à jour élaboré pour remplir ces objectifs est décrit au chapitre 4. Mais, d'abord, un nouveau plan de répartition géographique est décrit au chapitre 3.

3. VALEURS DE PEANO

La valeur de Peano est un paramètre qui permet de définir une certaine courbe de remplissage fractale. Cette valeur établit une transformation de R^2 dans R^1 de sorte que les points dans R^2 ou les objets dans l'espace soient arrangés dans un ordre unique (ordre de Peano) sur une liste. Dans la présente application, les objets sont les unités d'échantillonnage et l'espace R^2 est représenté par le système de coordonnées géographiques de la terre.

La valeur de Peano est obtenue par entrecroisement de nombres binaires (voir Peano (1908), Laurini (1987) et Saalfeld, Fifield, Broome et Meixler (1988). Soit $X = X_k \ldots X_3\, X_2\, X_1$ et $Y = Y_k \ldots Y_3\, Y_2\, Y_1$ la longitude et la latitude d'un point arbitraire en format binaire de k chiffres. Alors la valeur de Peano correspondante sera ou à l'extérieur de l'échantillon. Si un système de mise à jour ne respecte pas cette règle, les estimateurs de totaux et des autres paramètres de la population risquent d'être biaisés. Par exemple, considérons deux règles qui peuvent être utilisées dans le cas ii) pour l'échantillonnage de nouvelles sociétés crées par transfert. Une des possibilités est d'inclure les nouvelles sociétés dans l'échantillon si les sociétés mères faisaient partie de l'échantillon. Si ce n'est pas le cas, la nouvelle société peut faire l'objet d'un nouvel échantillonnage. Ainsi, la probabilité de sélection des nouvelles sociétés est multiple et peut conduire à une estimation biaisée, à moins que des ajustements soient effectués (ces ajustements se rapportent aux règles de multiplicité étudiées par Monroe Sirken (1970) et autres). La seconde possibilité est d'inclure les nouvelles sociétés dans l'échantillon si et uniquement si les sociétés mères faisaient partie de l'échantillon. Comme cette seconde règle traite de façon symétrique les changements dans l'uivers du sondage qui se produisent dans l'échantillon et à l'extérieur de celui-ci, l'estimation du paramètre sondé n'est pas biaisée.

Pendant l'élaboration d'un système de mise à jour, le statisticien ne doit pas se baser uniquement sur les propriétés statistiques des estimateurs, mais aussi sur le coût, la faisabilité et l'acceptation des règles par le client. Certaines règles peuvent exiger une collecte de données supplémentaires, entraînant par le fait même des coûts additionnels qui doivent être calculés dès la mise sur pied d'un nouveau sondage successif. Dans certains cas, les données additionnelles doivent être recueillis après la fin de l'étude. Cela peut s'avérer difficile, ou à tout le moins, peut entraîner des erreurs considérables qui ne sont pas dues à l'échantillonnage, et ainsi risquer d'introduire un biais. Certaines règles peuvent être applicables et peu coûteuses, sans toutefois satisfaire aux exigences du client ou des utilisateurs des données.

Le problème de la mise à jour des échantillons n'est pas un problème nouveau; des systèmes de mise à jour des échantillons sont utilisés depuis plusieurs années dans plusieurs des enquêtes successives de Statistique Canada, du United States Bureau of the Census et de la A.C. Nielsen Company. Néanmoins, il existe peu de littérature sur le sujet. Pour une brève discussion des problèmes de mise à jour, voir Wolter et coll. (1976) pour le cas ii), Hanson (1978) pour le cas iii) et Enrst (1989) pour le cas iv).

Dans la suite du présent article, nous étudierons principalement le cas i) où l'établissement est à la fois l'unité d'échantillonnage et l'unité élémentaire. C'est ce qui se produit dans nos sondages sur les établissements à la A.C. Nielsen Company. Le chapitre 2 décrit un de nos principaux sondages, le sondage Scantrack et les problèmes de mise à jour auxquels nous sommes confrontés dans ce sondage. Nous décrirons aussi certains des principaux objectifs que nous avions en élaborant un nouveau système de mise à jour pour ce sondage.

Le nouveau système de mise à jour est basé sur un paramètre connu en mathématique comme les valeurs de Peano, qui permettent de créer une courbe de remplissage fractale. Les valeurs de Peano sont définies dans le chapitre trois qui comprend aussi plusieurs graphiques pour faciliter la compréhension. La conclusion est donnée au chapitre 4 et comprend la description des règles du nouveau système de mise à jour de l'échantillon.

2. SONDAGE SCANTRACK

Les sociétés Nielsen fournissent des données provenant de plusieurs études de marché. L'unité d'échantillonnage des sondages portant sur les médias, comme le Nielsen Television Index et le Nielsen Station Index, est soit le logement ou le ménage. Pour les sondages portant sur l'industrie des biens de consommation comme le Nielsen Food Index, Le Nielsen Drug Index et le Nielsen Scantrack United States (NSUS), ce sont les magasins qui sont les unités d'échantillonnage. Le Single Source service, qui étudie les habitudes de consommation en fonction de l'écoute de la télévision et de la commercialisation des produits se sert du ménage et des magasins comme unités d'échantillonnage. Bien que la mise à jour de l'échantillon soit importante pour chacune de ces enquêtes, la présente discussion porte sur l'échantillon Scantrack de supermarchés à la base du service NSUS.

L'échantillon Scantrack compte 3 000 supermarchés répartis en 51 strates, 50 pour les grandes villes et une pour le reste des États-Unis. À l'intérieur d'une même strate ou marché, l'échantillon est stratifié à nouveau en fonction des grandes chaînes de supermarchés. La base est ordonnée géographiquement et un échantillon systématique est choisi à l'intérieur de chaque strate afin de bien représenter la situation socio-économique. Cet échantillon est aussi représentatif de l'âge du supermarché, de sa taille et des autres facteurs qui influent sur les ventes. Bien qu'un échantillon systématique ordonné géographiquement soit très simple et direct, le choix de ce plan d'échantillonnage est justifié par des années d'expérience et les résultats d'études empiriques portant sur l'essai de divers plans d'échantillonnage dans des conditions réelles.

Les supermarchés qui composent l'échantillon Scantrack sont munis de lecteurs optiques qui lisent les codes à barres sur les emballages au moment de l'achat. Les codes à barres sont appelés codes universels de produits ou CUP. Lorsque l'étiquette du produit est lue, la transaction est enregistrée dans l'ordinateur du magasin et le prix correspondant au code CUP est enregistré. Chaque semaine, le commerce nous indique le total des ventes et le prix de chaque article vendu. Comme le nombre d'articles portant un code CUP peut aller jusqu'à 10 000 dans un seul supermarché, nous traitons plus de 30 millions d'observations par semaine.

En plus des données des lecteurs optiques, nous recueillons les données portant sur la commercialisation d'un produit, que ce soit des coupons-rabais, la publicité parue dans un journal ou la disposition du produit dans le magasin. Lorsqu'un produit a fait l'objet d'une publicité, le type de publicité imprimée utilisée et l'emplacement de l'étalage dans le magasin sont aussi connus.

- Démolition d'un établissement existant.
- Établissement occupé de façon irrégulière.
- Changement de configuration d'un établissement, par exemple, s'il est divisé en deux ou plusieurs établissements.

Le cas ii) est beaucoup plus complexe que le cas i) puisque les unités d'échantillonnage sont des groupes d'unités élémentaires. Toutes les situations qui se produisent dans le cas i) peuvent aussi se produire dans le cas des sociétés constituées d'un seul établissement. Pour les sociétés formées de plusieurs établissements, les problèmes suivants s'ajoutent:

- Fusion de deux sociétés pour constituer une nouvelle société.
- Absorption d'une société par une autre. La société absorbante est la seule société résultante.
- Entreprise conjointe où deux sociétés collaborent pour former une nouvelle société qui peut être une filiale des deux sociétés mères.
- Opération de transfert où une société crée une nouvelle société indépendante.
- Opération de transfert où une société vend une partie de ses parts à une autre société.

Les situations qui se produisent dans le cas iii) ressemblent beaucoup à celles qui se produisent dans le cas i):

- Habitation nouvellement construite.
- Habitation qui passe d'une catégorie non observée à une catégorie observée.
- Habitation qui passe d'une catégorle observée à une autre catégorie observée.
- Habitation qui passe d'une catégorie observée à une catégorie non observée.
- Changement de fonction, de résidentielle à commerciale.
- Changement de fonction, de commerciale à résidentielle.
- Démolition d'habitations existantes.
- Reconfiguration d'édifices existants, par exemple, reconfiguration des appartements dans un petit édifice à plusieurs logements.

Enfin, le cas iv) ressemble beaucoup au cas ii) en terme de composition de la population de l'enquête et de la complexité des changements qui s'y produisent. Les problèmes de mise à jour sont les suivants:

- Mariage, création d'une nouvelle famille à partir de familles entières ou de parties de familles.
- Nouveaux membres qui entrent dans une famille existante, ce qui élimine une autre famille ou partie de famille.
- Divorces qui peuvent mener à la création d'une nouvelle famille à partir d'une famille existante.
- Déménagement des membres d'une famille, soit pour joindre une autre famille ou en créer une nouvelle.
- Naissance de nouveaux membres.
- Décès d'un membre.
- Déménagement d'une famille entière; il faut alors les retracer et peut-être modifier les charges de travail assignées.

Pour que l'échantillon reflète bien les changements qui se produisent dans l'univers de l'enquête et par le fait même demeure représentatif, l'organisation responsable du sondage doit élaborer un système de mise à jour explicite. Un système de mise à jour comprend un plan d'échantillonnage et une méthode de mise à jour de la base de sondage, possiblement énoncés, sous la forme de règles simples. Ces règles permettent au statisticien de bâtir l'échantillon de façon que la probabilité que chaque unité élémentaire soit incluse dans l'échantillon soit connue et différente de zéro pour toutes les périodes du sondage successif, ou, si c'est impossible, d'évaluer les données du sondage correctement afin d'en arriver à des estimateurs des paramètres d'intérêt non biaisés ou consistants. Dans les cas i à iv, il est évident que le système de mise à jour doit remplir au moins les quatre conditions suivantes:

- Donner aux nouvelles unités élémentaires une probabilité d'inclusion connue et différente de zéro.
- Tenir compte des unités élémentaires qui n'existent plus réellement.
- Empêcher que les unités élémentaires soient incluses plusieurs fois dans l'échantillon; si c'est impossible, le système doit tenir compte de la situation afin que les ajustements nécessaires puissent être faits au cours de l'estimation.
- Mettre la base de sondage à jour afin de faciliter et de contrôler les fonctions ci-dessus.

Tous les systèmes de mise à jour doivent respecter la règle suivante: le système ou les règles qui le définissent doivent traiter de façon symétrique les changements de l'univers de l'enquête, qu'ils se produisent à l'intérieur

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

MISE À JOUR DES ÉCHANTILLONS BASÉE SUR LES VALEURS DE PEANO

K.M. Wolter[1] et R.M. Harter[2]

RÉSUMÉ

Le présent document porte sur la mise à jour des échantillons et des bases utilisés dans les sondages successifs. Le système de mise à jour décrit remplit quatre objectifs principaux: 1) maintenir une répartition géographique équilibrée de l'échantillon, 2) maintenir constante la taille de l'échantillon, 3) maintenir le caractère non biaisé de l'estimateur et 4) empêcher l'apparition de distorsion dans l'estimation des tendances. Le système est basé sur les valeurs de Peano qui permettent de créer une courbe de remplissage fractale. L'exemple utilisé pour présenter le nouveau système est un sondage à l'échelle nationale portant sur les établissements des États-Unis, sondage effectué par la A.C. Nielsen Company.

1. INTRODUCTION

Dans le présent document, nous étudierons les sondages successifs et la mise à jour qu'ils exigent. Soit ν_t l'univers d'un sondage au temps t, t = 0 signifiant le début d'un nouveau sondage. Supposons aussi qu'un échantillon aléatoire d'unités de ν_0 a été choisi et qu'il est donc possible de construire des estimateurs non biaisés (ou du moins consistants) de la population totale et des autres paramètres d'intérêt. Cet univers est sondé régulièrement dans le temps, afin de déterminer le "niveau" de la population et de mesurer ses tendances. Un panel ou un plan de sondage avec renouvellement est habituellement utilisé à cette fin (voir Rao et Graham (1964) et Wolter (1979) et les documents de référence cités par ces auteurs). Dans de telles enquêtes portant sur les gens ou les institutions, les deux seuls points d'intérêt dans notre cas, la composition de la population varie avec le temps, en raison des naissances, des décès et des autres changements qui se produisent dans les unités d'échantillonnage. La base de sondage, le plan d'échantillonnage et les méthodes d'observation ou de collecte des données doivent être constants en fonction de tels changements; autrement, l'échantillon devient très biaisé et n'est plus représentatif de l'univers de l'enquête.

Le type de problèmes de mise à jour qui se produisent dans les sondages successifs dépend à la fois de l'univers étudié, du choix de l'unité d'échantillonnage et des interactions entre l'unité d'échantillonnage et les unités élémentaires de l'univers du sondage. Nous donnerons un bref résumé des problèmes qui se produisent dans quatre cas différents:

i) Enquête portant sur les établissements; unité d'échantillonnage: l'établissement.

ii) Enquête portant sur les établissements; unité d'échantillonnage: société ou autre regroupement similaire d'établissements.

iii) Enquête portant sur les gens ou les ménages; unité d'échantillonnage: adresse ou logement.

iv) Enquête sur les gens ou les ménages; unité d'échantillonnage: le ménage ou la famille.

Dans le cadre de ce texte. les mots "établissement" et "société" sont utilisés dans leur sens général. Un établissement peut être un commerce de détail, une usine de fabrication, une école, un hôpital, un terrain de golf ou toute autre entité localisée, tandis que la société est l'entité légale propriétaire du commerce, ou de la commission scolaire et ainsi de suite. Dans certains cas, l'établissement et la société sont une seule et même chose, par exemple, dans le cas d'un magasin d'alimentation indépendant.

Dans le cas i), l'univers de l'enquête peut varier comme suit:

- Établissement nouvellement construit.
- Établissement qui passe d'une catégorie non observée à une catégorie observée.
- Établissement qui passe d'une catégorie observée à une autre catégorie observée.
- Établissement qui passe d'une catégorie observée à une catégorie non observée.
- Changement de fonction d'un édifice, d'une fonction résidentielle à une fonction commerciale.
- Changement de fonction d'un édifice, d'une fonction commerciale à une fonction résidentielle.

[1] Kirk M. Wolter, Vice President, A.C. Nielsen, Northbrook, Illinois, 60062
[2] Rachel M. Harter, A.C. Nielsen, Northbrook, Illinois, 60062

Les échantillons symétriques périodiques (hebdomadaires, mensuels ou trimestriels) sont peut-être les plus simples et les meilleurs, mais l'utilisation d'une méthodologie autre peut être tolérée et même compensée au moyen de facteurs de pondération. En outre, il est possible, au besoin, d'ajouter des variables au contenu de base des enquêtes.

Bien que l'accent ait été mis sur les deux extrêmes (enquêtes ponctuelles pour l'actualité des données et cumuls complets pour tout l'intervalle (dix ans?) pour les petits domaines), il serait souvent souhaitable et faisable de procéder à des cumuls intermédiaires pour les domaines majeurs (provinces?) et les domaines mineurs (districts?). Il convient ici de préciser que lorsqu'il était question, implicitement ou explicitement, d'un recensement intégral complet, il pouvait également s'agir à la base de fractions élevées (10%) étant donné que le recensement couvrait tout l'intervalle, particulièrement là où des recensements décennaux sont également réalisés. De même, le terme "population" utilisé dans la définition peut de toute évidence englober plusieurs populations de même qu'un compte national de personnes.

Il faudrait examiner en bloc la taille et le facteur de pondération des échantillons périodiques; de plus, la plus grande souplesse est recommandée. Des travaux de recherche méthodologique peuvent être d'une grande utilité. Actuellement, nous supposons que les fractions de sondage sont identiques pour toutes les périodes et que seuls les facteurs de pondération de chacune des dix années sont différents, de sorte que le facteur de pondération total se chiffre à 10 pour la période de dix ans. Dans le cas de l'échantillon national et de variables très fluctuantes (par exemple, les maladies infectieuses), la dernièrr année peut porter le facteur de pondération complet de 10. Au contraire, pour les populations totales des petits domaines, chacune des dix années peut porter un facteur de pondération de 1. Cependant, pour de nombreuses variables et pour les grands domaines, il peut être préférable d'utiliser une moyenne mobile intermédiaire plutôt que l'un ou l'autre de ces extrêmes [par exemple : (100, 90, 80, 65, 50, 40, 30, 20, 15, 10)/50].

J'aimerais terminer par une citation (Kish, 1986) ayant trait aux échantillons successifs et à d'autres enquêtes-échantillon périodiques (ou répétées), "Cinquièmement, la stratégie statistique devrait supposer une déclaration moins fréquente, surtout pour les domaines petits mais non négligeables. Il arrive trop souvent que de tels domaines ne fassent pas l'objet d'enquêtes, ou que celles-ci comportent de trop grandes erreurs ou qu'elles soient trop coûteuses, ou les deux à la fois. Citons comme exemple les vastes régions peu peuplées de nombreux pays et pour lesquelles les autorités provinciales exigent des rapports distincts. Entre autres exemples, notons également les groupes démographiques, ethniques, professionnels, etc. pour lesquels il faut des données distinctes. Plutôt que de perpétuer la pratique rigide habituelle qui prévaut actuellement, il serait préférable d'accroitre la fréquence de déclaration pour ces petits domaines, d'établir un plan de sondage pour des périodes plus longues et de procéder au cumul des échantillons. Les tableaux présentant les statistiques ainsi obtenues devraient indiquer les différents plans de sondage utilisés."

BIBLIOGRAPHIE

Kish, L. 1965. Survey Sampling, New York and Sydney: John Wiley and Sons.

Kish, L. 1979. Rotating Samples Instead of Censuses, Census Forum, Honolulu: East-West Center 6, 1-13.

Kish, L. 1981. Using Cumulated Rolling Samples, Washington: Library of Congress, U.S. Government Printing Office, No 80-528.

Kish, L. 1983. Data Collection over Time and Space, in T. Wright, Statistical Methods and the Improvement of Data Quality, Orlando: Academic Press, 73-84.

Kish, L. 1986. Timing of Surveys for Public Policy, Australian Journal of Statistics, 28(1), 1-12.

Kish, L., and Verma, V. 1986. Complete Censuses and Samples, Journal of Official Statistics, 2, 381-94.

Kish, L. 1987. Statistical Design for Research, New York: John Wiley and Sons, Section 6.5.

Kish, L. 1989. Sampling Methods for Agricultural Surveys, Rome: FAO, Statistics Division, Section 16.3.

Les échantillons successifs présentent aussi un problème, moins grave bien qu'ennuyeux, attribuable aux déplacements (des personnes, des ménages, etc.), de sorte que les mêmes unités peuvent se trouver dans deux ou même plusieurs échantillons périodiques. Le choix arbitraire d'une date du recensement a pour effet d'annuler l'incidence de ces déplacements, bien que l'application de cette méthode soit coûteuse, arbitraire et erronée. De tels déplacements se produisent aussi dans le cadre d'enquêtes ponctuelles. Ils se compteront toutefois par milliers dans les échantillons successifs cumulés. Cependant, pour ce qui est de la sélection aléatoire de segments de régions pour établir des échantillons probabilistes, les déplacements ne causent aucun biais. Nous avons seulement besoin de comprendre et d'expliquer.

Les problèmes de coûts liés à un recensement complet avec échantillons successifs semblent énormes en regard des coûts de la plupart des enquêtes périodiques. L'écart est cependant moins marqué dans les petits pays parce que les fractions de sondage y sont plus élevées. Par exemple, une enquête mensuelle sur la population active menée auprès de 80,000 ménages exige une fraction de sondage de seulement 1/1000 dans un grand pays comptant 80 millions de ménages alors que cette dernière s'élève à 1/100 dans un pays comprenant 8 millions de ménages; ces échantillons s'additionneraient sans chevauchement de façon à produire un recensement complet à l'issue des 120 mois de la période de dix ans. Nous aborderons bientôt la question des chevauchements. En outre, le coût par interview dans le cas des échantillons successifs est nécessairement plus élevé que celui des échantillons actuels qui sont confinés aux régions primaires d'échantillonnage. Cependant, les coûts de déplacement seraient loin d'augmenter autant que ce que laisserait supposer la présence de petites régions d'UPÉ sur les cartes du territoire national. Dans chaque pays, la majorité des unités de l'échantillon et de la population sont concentrées dans un nombre relativement petit de "régions autoreprésentatives".

Cependant, pour calculer les coûts légitimes, nous devons additionner les coûts des enquêtes sur la population active et ceux des recensements décennaux et peut- être quiquennaux, puisque les échantillons successifs sont censés faire le travail des deux. Il est peut-être vrai que les employés affectés au recensement sont d'ordinaire peu rémunérés, mais les coûts de recrutement et de formation pour seulement quelques jours de travail peuvent être relativement élevés.

Le problème de la qualité de la couverture de certains recensements comparativement à celle des enquêtes-échantillon, comme il en a déjà été fait mention, est trop technique et trop précis pour n'être traité que brièvement ici. Il est probable que des efforts spéciaux permettraient d'améliorer la couverture des enquêtes-échantillon. Par exemple, le USCB s'efforce de vérifier et d'améliorer le recensement de 1990 à l'aide d'une enquête-échantillon spéciale auprès de 150,000 ménages.

Dans le cadre des enquêtes périodiques sur la population active et de certaines autres enquêtes, on a souvent recours à des chevauchements considérables pour deux principales raisons. La raison la plus importante est la moins souvent mentionnée : le coût des interviews tardives est moins élevé que celui des premières interviews, surtout lorsqu'elles ont lieu par téléphone. Bien que cet écart ne soit pas très élevé, il demande considération lors de comparaisons. Les raisons les mieux connues sont celles qui s'expriment à l'aide de formules de corrélations entre les plans de réinterviews; ces dernières réduisent légèrement les variances dans le cas des estimations courantes, et davantage lorsqu'il s'agit de l'estimation des variations nettes (ou macro-variations) entre les périodes. Toutefois, ces corrélations sont faibles pour de nombreuses variables d'enquête, comme les mesures du chômage. Elles sont davantage affaiblies par les erreurs de réponse et par les taux mobiles qui approchent de 0.2 entre les années.

Ainsi, les corrélations sont plus faibles lorsque le chevauchement implique des segments plutôt que des personnes; mais de tels chevauchements sont plus faciles à manier, moins coûteux et libres des biais attribuables aux panels de personnes. En revanche, un panel de personnes se traduirait par des corrélations plus élevées et permettrait aussi l'analyse des variations individuelles, c'est-à- dire les micro-variations ou variations brutes. En raison de cette contradiction, certaines enquêtes ont fait les deux : couvrir les mêmes segments et suivre aussi les personnes qui se déplacent pour maintenir les panels.

La taille et la nature du chevauchement des échantillons commandent la réalisation d'études techniques; ces études devraient être à usages multiples parce que les corrélations varieront grandement d'une variable à l'autre. Le conseil que je formule ici sans façon a trait aux chevauchements qui représenteraient le tiers ou moins de la portion cumulée sans chevauchement. De plus, le chevauchement pourrait représenter un panel de personnes suivies pendant de nombreuses périodes, permettant ainsi une analyse dynamique des variations individuelles, qu'il est actuellement impossible de faire à partir des enquêtes sur la population active. Les chevauchements s'étendant sur de nombreuses périodes réduiraient les écarts entre les variations nettes pour toutes les paires de périodes. Je propose d'attribuer le nom de "plans à panel fractionné" (PPF) à de tels plans de sondage [Kish, 1982, 1986, 1987].

Le plan de base met en jeu F sondages périodiques permettant d'établir des estimations fréquentes de la population (hebdomadaires, mensuelles ou annuelles) et comporte le cumul, pour l'ensemble de l'intervalle, des F échantillons à l'origine des estimations pour les petits domaines. Cette définition peut laisser une grande liberté d'action. Voici maintenant quelques exemples. D'abord, il est possible d'établir de meilleures estimations pour les domaines (provinces) à l'aide de fractions de sondage plus élevées et de cumuls s'étendant sur de plus longues périodes. Avec des estimations trimestrielles plutôt que mensuelles et des taux de sondage triplés, la taille de l'échantillon initial serait multipliée par un facteur de 9, par exemple.

intervalles fréquents (hebdomadaires, annuels). Aujourd'hui, la publicité favorise les comptes détaillés de la population pour les régions administratives, mais les données détaillées pour d'autres domaines, pour des "domaines de recoupement" (comme l'âge et les classes sociales) peuvent se révéler tout aussi importantes à long terme.

Nous devons mentionner en premier lieu les recensements décennaux de la population, des logements, de l'agriculture, de l'industrie et autres, que l'humanité a étendu sur toute la surface de la terre au cours des deux derniers siècles, et surtout au cours des deux dernières décennies, avec l'aide des Nations Unies. Outre les données détaillées relatives aux petits domaines, les recensements peuvent aussi parfois permettre d'obtenir une meilleure couverture en raison de la publicité concentrée et du "rituel" national entourant la tenue des recensements. Le recensement chinois de 1982 constitue un bon exemple. La concentration des efforts déployés dans le cadre d'un recensement peut également se traduire par des coûts unitaires inférieurs à ceux des enquêtes; cependant, le recensement des États-Unis de 1990, qui nécessitera un financement de 2.6 milliards de dollars, coûtera $10 par habitant ou $30 par ménage. Toutefois, l'utilisation d'échantillons successifs est proposée ici surtout parce que les données des recensements décennaux ne sont pas assez actuelles : entre la collecte et l'utilisation des données, il s'écoule généralement une période de 1 à 14 ans. [Kish, 1981].

Il a aussi été proposé d'accroître la fréquence des recensements, c'est-à-dire de réaliser un recensement tous les cinq ans ou chaque année. Cependant, les recensements quinquennaux ne seraient pas assez fréquents et les recensements annuels seraient trop coûteux. La réalisation de recensements-échantillon de 1% ou 10% de la population a été proposée, mais l'échantillon serait trop petit dans le premier cas et les coûts seraient trop élevés dans le deuxième. Dans au moins deux pays, des recensements quinquennaux de 10% de la population ont été réalisés pour la moitié des coûts d'un recensement complet; de plus, dans un des cas, on a observé une diminution de la couverture. Le micro-recensement de 1% de la population en Allemagne de l'Ouest et les échantillons 1/2000 de la Chine fournissent certaines données annuelles. Le Canada a réalisé un recensement de 10% de la population en 1985. Je doute que ces efforts permettront de répondre en général, aux besoins en données tant actuelles que détaillées. Pour paraphraser une parole de Lincoln : "On ne peut pas recueillir des données auprès de tout le monde, tout le temps."

Cette phrase nous amène à examiner le recours aux registres et dossiers administratifs comme méthode de collecte de données à la fois actuelles et détaillées. Comme exemples frappants, mentionnons les registres de la population des pays nordiques : la Suède, la Norvège, le Danemark et la Finlande, et peut-être quelques autres pays. Dans certains cas, ils ont remplacé les données de recensement par des données tirées de registres, ou ils sont susceptibles de le faire. L'intégralité de ces données dépend de la collaboration, de la motivation et de l'alphabétisation. Dans d'autres cas, la couverture, la qualité et la mise à jour des registres sont loin d'être adéquates. Je prévois un accroissement de leur qualité, de leur portée et de leur utilisation, mais je ne crois pas que ces registres remplaceront les recensements ni bientôt ni complètement, parce que leur contenu est susceptible de se limiter à quelques variables de base, trop peu nombreuses pour répondre aux besoins d'aujourd'hui en données du recensement. Pour paraphraser encore une fois une parole de Lincoln : "On ne peut pas recueillir des données auprès de tout le monde, tout le temps, sur tous les sujets."

Qu'en est-il de l'estimateur synthétique et de l'estimateur de la méthode itérative du quotient qui fourniraient des estimations actuelles et détaillées fondées sur des données de recensement, plus les registres et les enquêtes? Je suis optimiste quant aux progrès que ces méthodes réaliseront, mais je ne crois pas qu'elles en viennent à remplacer la collecte de données au moyen de recensements ou d'échantillons successifs.

Nous devons maintenant discuter de trois problèmes soulevés par les échantillons successifs cumulés : leurs coûts, leur couverture et leurs bases sur lesquelles les moyennes de populations en mutation sont calculées. Ces moyennes doivent avoir trait tant aux variations démographiques dans le temps qu'aux variations individuelles dans l'espace.

Pour faire la moyenne des variations dans le temps, il faut surmonter des blocages psychologiques issus de la tradition et de la pratique concernant les recensements et les enquêtes. J'ai déployé beaucoup d'efforts pour surmonter ces blocages en invoquant des arguments fondés sur l'inférence et la philosophie statistiques; de plus, nous avons besoin d'un plus grand nombre de travaux théoriques, méthodologiques et empiriques. Le cumul d'enquêtes répétées pendant un intervalle de temps complet peut se traduire par une meilleure inférence statistique qu'une seule enquête ponctuelle. La sélection aléatoire de segments de temps dans un intervalle entier permet, grâce à une inférence statistique à partir de l'échantillon, d'arriver à une tendance moyenne pour l'intervalle. Au contraire, l'inférence à partir du segment de temps "typique" d'une enquête ponctuelle afin d'obtenir des données pour tout l'intervalle exige le recours à des jugements de valeur, à des hypothèses et à des modèles en ce qui concerne la nature de la variation, ou du manque de variation, pour l'intervalle entier. Le choix d'un seul segment de temps s'accompagne des risques de variations saisonnières, cycliques, séculaires et ponctuelles, connues ou non. Le cumul d'enquêtes répétées repose sur la moyenne des variations observées lors des enquêtes répétées [Kish, 1965, 12.5D]. Sur le plan statistique et méthodologique, l'échantillonnage et le cumul dans le temps devraient être préférables à l'acceptation d'une période "typique" choisie arbitrairement. Il est paradoxal que le choix arbitraire soit encore accepté et pratiqué lorsqu'il s'agit de l'aspect temporel, alors que nous refusons de tolérer le choix arbitraire des segments spatiaux dans le cadre d'échantillonnage probabiliste [Kish, 1979, 1981, 1983, 1986].

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

CARACTÉRISTIQUES ET PROBLÈMES PROPRES AUX ÉCHANTILLONS SUCCESSIFS

L. Kish[1]

Je vous suis reconnaissant de me donner ainsi l'occasion d'expliquer les principales caractéristiques des échantillons successifs, ainsi que les utilisations que l'on peut en faire. D'abord, permettez-moi d'ébaucher une définition de la notion d'échantillons successifs. Il s'agit d'un plan de sondage combiné (commun) comportant F échantillons périodiques, chacun d'entre eux étant un échantillon aléatoire de l'ensemble de la population, conçu de telle sorte que le cumul des F périodes se traduit par un recensement détaillé de la population entière; de plus, le cumul de périodes intermédiaires devrait permettre l'obtention de renseignements détaillés pour une à F périodes. Nous pouvons évaluer l'exactitude de cette définition en examinant des exemples et des contre-exemples. Nous examinerons également les variations possibles pouvant répondre à la définition et les besoins contradictoires que les échantillons successifs peuvent viser à satisfaire.

Imaginez des échantillons nationaux hebdomadaires, chacun d'entre eux étant établi selon un taux d'échantillonnage équiprobable de 1/520, conçus de telle façon que l'ensemble de la population ait été échantillonnée au bout de 520 semaines et que le cumul se traduise en moyenne par un recensement complet de la population pour la période de dix ans. Pour chaque année, on obtiendrait des échantillons nationaux et locaux dont le taux d'échantillonnage serait de 52/520, soit 1/10. Le plan de sondage combinerait des échantillons nationaux hebdomadaires avec en moyenne un recensement décennal complet, et avec des recensements-échantillon de 10% de la population chaque année.

J'utilise le conditionnel parce que le plan de sondage en question n'existe encore nulle part, pour autant que je sache. J'ose inventer cette définition, étant donné que je l'ai utilisée pour la première fois en 1981 dans un rapport publié destiné à un comité du Congrès américain [Kish, 1981]. J'ai déjà décrit de tels plans dans un document intitulé "Rotating samples instead of censuses" (Échantillons avec renouvellement au lieu de recensements)[Kish, 1979]. Le terme "échantillons avec renouvellement" a cependant suscité des objections parce qu'il entraînait une certaine confusion avec les échantillons à chevauchements partiels qui sont largement utilisés dans le cadre des enquêtes sur la population active. En inventant le terme "échantillons successifs", je veux éviter une confusion inutile avec d'autres plans de sondage. En attribuant un nom descriptif à mes méthodes, j'espère également favoriser une meilleure compréhension. L'utilisation de noms descriptifs permet d'éviter que les méthodes soient désignées par le nom de leur auteur, pratique agaçante qui entraîne un antagonisme inutile au sujet des priorités.

Les enquêtes sur la population active qui sont réalisées actuellement, telles que la CPS aux États-Unis et l'EPA au Canada, diffèrent grandement des enquêtes mettant en jeu des échantillons successifs. Premièrement, les enquêtes sur la population active comportent généralement un nombre considérable de chevauchements, ce qui entrave et retarde le cumul des périodes. Deuxièmement, elles sont menées uniquement dans les régions primaires d'échantillonnage, de sorte que le cumul des données ne permet pas de couvrir l'ensemble du territoire national. Troisièmement, la taille des échantillons peut ne pas être assez grande pour permettre d'arriver par cumul à un recensement complet. Quatrièmement, les méthodes utilisées ont tendance à se traduire par une couverture moins complète que celle du recensement.

Les échantillons sans chevauchement sont parfois appelés sondages "échelonnés" et servent à cumuler des données ayant trait à de courtes périodes de rappel. Dans des pays en voie de développement, ils ont servi à recueillir des données démographiques, comme les taux de natalité et de mortalité. Les 52 échantillons hebdomadaires sans chevauchement de 1000 ménages établis dans le cadre de la HIS du NC HS peuvent constituer un bon exemple. Cependant, ces échantillons sont également trop petits et trop confinés aux UPÉ pour être considérés comme des échantillons successifs qui se traduiraient par un recensement national détaillé.

À la réunion de la ASA tenue en août 1989, une méthode consistant à alterner les années de recensement complet dans les États a été proposée sous la fausse appellation d'"échantillons successifs". J'ai demandé aux auteurs d'éviter cette confusion inutile. En outre, je crois que cette idée a peu de chances de succès. Cette méthode reviendrait à confondre les variations annuelles et les variations d'un État à l'autre, ce qui entraînerait une confusion des comparaisons temporelles et spatiales. Elle irait ainsi dans le sens contraire des efforts de l'ONU en vue d'établir des dates de collecte décennale. Nous pouvons arrêter là les critiques, mais retenez uniquement qu'il s'agit d'un contre-exemple au sujet des échantillons successifs.

Nous devons maintenant examiner très brièvement les principaux moyens autres que les échantillons successifs pour obtenir des renseignements détaillés relatifs aux petits domaines, qui est l'un des principaux objectifs du cumul d'échantillons successifs, l'autre étant de fournir des estimations globales de la population à des

[1] L. Kish, Institute for Social Research, University of Michigan, Ann Harbor, Michigan, U.S.A.

Gurney, M. et Daly, J.F. (1965), "A Multivariate Approach to Estimation in Periodic Sample Surveys," *Proceedings of the Social Statistics Section of the American Statistical Association*, 242-257.

Huang, E.T. et Fuller, W.A. (1978), "Nonnegative Regression Estimation for Sample Survey Data," *Proceedings of the Social Statistics Section of the American Statistical Association*, 300-303.

Jessen, R.J. (1942), "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

Jones, R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.

Kalton, G. (1983), *Compensating for Missing Survey Data*, University of Michigan, Survey Research Center.

Kasprzyk, D., Duncan, G.J., Kalton, G., et Singh, M.P. (1989), *Panel Surveys*, New York: John Wiley.

Kasprzyk, D. et McMillen, D.B. (1987), "SIPP: Characteristics of the 1984 Panel," *Proceedings of the Social Statistics Section of the American Statistical Association*, 181-186.

Lazarsfeld, P.F. et Fiske, M. (1983), "The Panel as a New Tool for Measuring Opinion," *Public Opinion Quarterly*, 2, 596-612.

Lepkowski, J.M. (1989), "Treatment of Wave Nonresponse in Panel Surveys," in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh, New York: John Wiley.

Little, R.J.A. et Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Little, R.J.A. et Su, H.L. (1989), *Item Nonresponse in Panel Surveys*, in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh, New York: John Wiley.

Madow, W.G., Olkin, I., Nisselson, H., et Rubin, D.B. (1983), *Incomplete Data in Sample Surveys*, (trois volumes) New York: Academic Press.

Patterson, H.D. (1950), "Sampling on Successive Occasions with Partial Replacement of Units," *Journal of the Royal Statistical Soc.*, B12, 241-255.

Poterba, J.M. et Summers, L.H. (1985), "Adjusting the Gross Change Data: Implications for Labor Market Dynamics," *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census et U.S. Bureau of Labor Statistics, pp. 81-95.

Rao, J.N.K. et Graham, J.E. (1964), "Rotation Designs for Sampling on Repeated Occasions", *Journal of the American Statistical Association*, 59, 492-509.

Smith, T.M.F. et Holt, D. (1989), "Some Inferential Problems in the Analysis of Surveys Over Time," article présenté à la 47ième sessions de International Statistical Institute, Paris.

Wolter, K. (1979), "Composite Estimation in Finite Populations," *Journal of the American Statistical Association*, 74, 604-613.

$$\hat{\gamma} = (\hat{\theta}_{11} - \hat{\theta}_1^2)^{-1} (\hat{\theta}_1 - \hat{\theta}_1^2)$$

et

$$\hat{P}_{11} = \hat{\gamma} (\tilde{P}_{11} - \tilde{P}_{1.} \tilde{P}_{.1}) + \tilde{P}_{1.} \tilde{P}_{.1},$$

où

$$\theta_1 = \theta_{11} + \theta_{12} = \theta_{11} + \theta_{21}$$

$\hat{\theta}(ij)$ sont les estimations découlant de la réinterview et $\tilde{P}(ij)$, les estimations découlant des interviews réalisés aux deux périodes.

Dans la construction de l'estimateur, les résultats de la réinterview ne servent qu'à estimer le paramètre de l'erreur de mesure. En fait, ces résultats pourraient être utilisés dans une méthode des moindres carrés généralisés pour améliorer les valeurs estimées de $P(11)$, $P(1.)$ et $P(.1)$. En supposant qu'il en coûte la même chose pour chaque interview, nous pouvons montrer qu'environ le quart des ressources devraient être consacrées à la réinterview. Le tableau 6 indique l'efficacité des estimateurs redressés en fonction de l'erreur de mesure par rapport aux estimateurs directs biaisés.

Tableau 6. Efficacité relative des estimateurs redressés en fonction de l'erreur de mesure

	Taille de l'échantillon (n)			
	500	1 000	5 000	10 000
EQM (direct)/EQM (erreur de mesure)	0.87	1.13	3.22	5.84

Pour de petits échantillons, l'erreur quadratique moyenne est moins élevée dans le cas de l'estimateur direct parce que la variance de cet estimateur est moins élevée. Rappelons-nous que seulement les trois quarts des observations nous renseignent sur $P(EE) = P(11)$. Toutefois, pour des échantillons dont la taille est supérieure à 750, le carré du biais de l'estimateur direct représente la très grande partie de l'erreur quadratique moyenne de cet estimateur, laquelle est plus élevée que celle de l'estimateur redressé en fonction de l'erreur de mesure. Ce court exemple illustre bien l'efficacité des plans de sondage qui permettent d'estimer les paramètres du processus d'erreur de mesure.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à un contrat de coopération (n° 68-3A75-8-12) passé avec le Soil Conservation Service du Département de l'agriculture des É.-U. L'auteur remercie Margot Tollefson pour les calculs.

BIBLIOGRAPHIE

Abowd, J.M. et Zellner, A. (1985), "Estimating Gross Labor Force Flows," *Journal of Business and Economic Statistics*, 3, 254-283.

Battese, G.E., Hasabelnaby, N.A., et Fuller, W.A. (1989), "Estimation of Livestock Inventories Using Several Area - and Multiple-Frame Estimators," *Techniques d'enquêtes*, 15, 13-27.

Chua, T.C. et Fuller, W.A. (1987), "A Model for Multinomial Response Error Applied to Labor Flows," *Journal of the American Statistical Association*, 82, 46-51.

Cochran, W.G. (1942), "Sampling Theory When the Sampling Units are the Unequal Sizes," *Journal of the American Statistical Association*, 37, 199-212.

Duncan, G.J. et Kalton, G. (1987), "Issues of Design and Analysis of Surveys Across Time," *International Statistical Review*, 55, 97-117.

Eckler, A.R. (1955), "Rotation Sampling," *The Annals of Mathematical Statistics*, 26, 664-685.

Garcia, P.A., Battese, G.E., et Brewer, W.D. (1975), "Longitudinal Study of Age and Cohort Influences on Dietary Patterns," *Journal of Gerontology*, 30, 349-356.

Graham, J.E. (1973), "Composite Estimation in Two Cycle Rotation Sampling Designs," *Communications in Statistics*, 1, 419-431.

où α est le paramètre du mécanisme de réponse. Selon ce modèle, l'espérance de la proportion de chômeurs à n'importe quelle période est égale à la proportion réelle. Un estimateur convergent de P(EE) selon le modèle de Chua-Fuller est

$$\hat{\pi}_{EE} = (1 - \alpha)^{-2}\{\hat{p}_{EE} - \hat{p}_{E.}\ \hat{P}_{.E}\ [1 - (1 - \alpha)^{2}]\} \quad ,$$

où $\hat{P}(EE)$, $\hat{P}(E.)$ et $\hat{P}(.E)$ sont les estimateurs directs et α est un paramètre du mécanisme de réponse. Voir aussi Battese et Fuller (1973). Compte tenu des résultats de la réinterview, une valeur α = 0.10 n'est pas exagérée. Par conséquent, nous avons

$$\hat{\pi}_{EE} = (0.90)^{-2}\{0.91 - 0.93(0.94)(0.19)\}$$
$$= 0.9184.$$

Le tableau des proportions corrigées en fonction de l'erreur de réponse est

$$\begin{pmatrix} 0.9184 & 0.0116 \\ 0.0216 & 0.0484 \end{pmatrix} .$$

Dans cet exemple, le biais de l'estimateur direct de P(EE) est 0.0084. Chua et Fuller estiment le même biais à 0.0168 dans le tableau de contingence qui comprend aussi les inactifs. Dans le tableau 5, nous comparons des méthodes d'estimation pour P(EE). Nous supposons un échantillon de 10 000 unités. Les trois méthodes de la portion gauche du tableau sont celles du tableau 3. Les trois autres représentent les estimateurs redressés en fonction du biais dû à l'erreur de mesure. Dans le calcul de la variance, on suppose que α a une erreur type de 0.01. La correction effectuée pour tenir compte du biais dû à l'erreur de mesure ne modifie pas les estimateurs de P(E.) et de P(.E). Dans cet exemple, le carré du biais de l'estimateur ordinaire de P(EE) équivaut à environ neuf fois la variance de l'estimateur par les moindres carrés généralisés. Le biais dû à l'erreur de mesure explique donc en très grande partie l'erreur quadratique moyenne de l'estimateur de P(EE).

Tableau 5. Erreur quadratique moyenne de divers estimateurs pour un échantillon de 10 000 unités à chaque période et un taux de chevauchement de 50 % (erreur quadratique moyenne de l'estimateur par les MCG redressé en fonction de l'erreur de mesure = 100).

	Méthode d'estimation					
	Ordinaire			Erreur de mesure		
Paramètre	Simple	MCG avec contrainte	MCG	Simple	MCG avec contrainte	MCG
$P_{E.}$	111	111	100	111	111	100
$P_{.E}$	111	101	100	111	101	100
P_{EE}	1071	967	961	250	106	100

Ces résultats ont des conséquences importantes pour ce qui a trait à la conception du plan de sondage. Afin d'illustrer cela, revenons au problème de la variation brute. Supposons que nous voulions estimer la probabilité qu'une personne appartienne à la catégorie des personnes avec emploi pendant deux périodes données, P(EE). Nous supposons qu'il est possible de réaliser des réinterviews indépendantes aux deux périodes et que les interviews réalisées à deux périodes quelconques sont indépendantes. Nous supposons aussi qu'il n'y a que deux scénarios d'interview possibles :

A. Interview et réinterview à l'une des deux périodes.

B. Interview à la période 1 et interview à la période 2.

Enfin, nous supposons que l'erreur de réponse est non biaisée et qu'un modèle à deux catégories (personnes avec emploi et chômeurs) est approprié dans les circonstances. Nous supposons aussi que la probabilité qu'une réponse soit exacte dépend uniquement de la catégorie à laquelle appartient le répondant dans la période courante.

Soient les probabilités de réponse définies en fonction de α et soit

$$\gamma = (1 - \alpha)^{-2} .$$

Désignons par θ(ij) l'élément ij de la matrice de probabilités 2 x 2 observée par suite de la réinterview. Par conséquent, θ(ij) est la probabilité qu'une personne réponde i à l'interview et j à la réinterview. Pour ce modèle simple, il existe des formules explicites pour les estimateurs. Ainsi, Les autres chiffres du tableau sont essentiellement des estimations obtenues par les moindres carrés, qui satisfont les six totaux de contrôle. Au cours du processus d'estimation, on a dû parfois recourir à des méthodes d'imputation, par exemple lorsqu'un changement observé dans les données des segments ne se reflétait pas dans les données des points d'observation.

Tableau 4. Illustration de la méthode d'estimation

1982	1987				
	Terrain labourable	Autres	Terrain urbain	Routes	TOTAL
Terrain labourable	26,243	179	13	6	26,441
Autres	771	7,114	6	2	7,893
Terrain urbain	0	0	623	0	623
Routes	17	4	0	1,038	1,059
TOTAL pour 1987	27,031	7,297	642	1,046	36,016

Le plan de l'enquête s'est traduit par des variances élevées pour les estimations directes de la variation de superficie dans le cas des modes d'utilisation du sol relativement moins importants comme le terrain urbain, le terrain labourable et les petites nappes d'eau. On s'est donc servi d'une méthode d'estimation pour petites régions pour établir des estimations de variation de superficie pour les zones principales de sols comprises dans les limites d'un comté (MLRAC). A cette fin, on a utilisé un programme d'ordinateur élaboré à l'Université Iowa State. Fuller (1986) expose la théorie qui est à la base de la méthode d'estimation pour petites régions. Le programme d'ordinateur a permis d'établir des estimations de la variation de superficie pour cinq modes d'utilisation du sol secondaires pour chacune des 5 500 MLRAC. Il s'agit là essentiellement d'une opération de répartition en ce sens que la somme des estimations pour les MLRAC équivaut à l'estimation globale pour l'État. Par ailleurs, on a établi des estimations pour les éléments du tableau 4 (élargi à 14 modes d'utilisation du sol) pour chaque MLRAC. A cette occasion, les estimations régionales pour les MLRAC, les estimations de la superficie représentée par les routes et les estimations de la superficie totale de terrain labourable pour l'État ont servi de totaux de contrôle. Le processus d'estimation s'est terminé par la pondération des données des points d'observation de manière à obtenir les estimations du tableau 4 pour chaque MLRAC.

En résumé, le processus d'estimation aboutit à une série de données de totalisation qui se rapportent à des points d'observation et qui permettent d'estimer tous les éléments d'un tableau à double entrée décrivant, pour n'importe quelle région identifiable, l'évolution de la superficie associée à divers modes d'utilisation du sol pour la période 1982-1987. Les estimations ainsi obtenues concordent avec les estimations établies antérieurement à l'échelle de l'État pour les principaux modes d'utilisation du sol et concordent aussi avec les données provenant de sources autres que les points d'observation.

En règle générale, l'échantillon de totalisation ne produit pas de bonnes estimations de la variance même si les segments et les strates sont bien identifiés dans la série de données. A cause du contrôle qui a été exercé sur l'échantillon de 1982, les données des points d'observation relatives aux principaux modes d'utilisation du sol, comme le terrain labourable, produiront des estimations de la variance trop élevées.

5. ERREUR DE MESURE

L'erreur de mesure peut avoir une incidence notable sur l'analyse des données dans le temps. Cette incidence peut être modérée dans le cas de moyennes observées périodiquement mais peut aussi être très appréciable dans le cas de l'estimation de la variation brute ou de l'estimation par régression.

Pour montrer jusqu'à quel point l'erreur de mesure peut biaiser les estimateurs de la variation brute, nous allons reprendre l'exemple du tableau 1. Chua et Fuller (1987) montrent que les proportions estimées qui figurent dans les cases du tableau à double entrée seront fortement biaisées si les données sont recueillies au moyen de la méthode utilisée par le U.S. Bureau of the Census. Voir aussi à ce sujet Abowd et Zellner (1985) et Poterba et Summers (1985). Le modèle de Chua-Fuller suppose que les erreurs de réponse aux deux périodes sont indépendantes. Il suppose aussi que, pour chaque période,

$$P\,(\text{réponse} = E \mid \text{situation réelle} = E) = 1 - \alpha + \alpha P_E,$$

$$P\,(\text{réponse} = U \mid \text{situation réelle} = E) = \alpha P_U,$$

$$P\,(\text{réponse} = U \mid \text{situation réelle} = U) = 1 - \alpha + \alpha P_U,$$

$$P\,(\text{réponse} = E \mid \text{situation réelle} = U) = \alpha P_E,$$

L'estimateur que nous venons de décrire sera efficace dans la plupart des circonstances. Cependant, il se peut qu'il produise des valeurs estimées négatives pour des quantités réputées non négatives parce qu'il est linéaire et que certains poids peuvent être négatifs. Des méthodes ont été mises au point pour corriger cette lacune. Voir Huang et Fuller (1978).

4. INVENTAIRE DES RESSOURCES NATIONALES DES É.-U.

Le Iowa State Statistical Laboratory collabore avec le U.S. Soil Conservation Service à la réalisation d'une enquête d'envergure sur l'utilisation du sol aux États-Unis. Des enquêtes ont déjà eu lieu en 1958, 1967, 1975, 1977, 1982 et 1987. On en prévoit une autre en 1992.

Cette enquête permet de recueillir des données sur la nature et l'utilisation du sol, le couvert végétal, la possibilité de transformer des terres qui ne servent pas actuellement à la culture en terres labourables, l'érosion hydrique et les méthodes de conservation. La collecte des données est confiée à des employés du U.S. Soil Conservation Service tandis que l'université Iowa State s'occupe de l'élaboration du plan de sondage et de l'estimation.

L'échantillon est un échantillon stratifié des terres non fédérales de 49 États (l'Alaska étant exclu) et de Porto Rico. Les unités d'échantillonnage sont des portions de terrain appelées segments. La superficie de ces segments varie de 40 à 640 acres. Des données sont recueillies pour tout le segment en ce qui concerne des aspects comme le sol urbain et les plans d'eau. En revanche, des données détaillées sur la nature et l'utilisation du sol sont recueillies à certains endroits dans le segment, choisis aléatoirement. En règle générale, on compte trois points d'observation par segment; les segments de 40 acres n'en comptent que deux et les segments des échantillons de deux États n'en comptent qu'un. Certaines données, comme la superficie totale et la superficie représentée par les routes, sont recueillies au moyen d'un recensement qui n'a rien à voir avec l'enquête précitée.

En 1982, l'échantillon comprenait environ 350 000 segments et près d'un million de points d'observation. En 1987, il comptait environ 100 000 segments, dont la majeure partie provenait de l'échantillon de 1982. Néanmoins, environ 1 500 nouveaux segments, prélevés dans des régions à forte croissance urbaine, ont été inclus dans l'échantillon de 1987. De plus, celui-ci comptait environ 280 000 points d'observation.

L'enquête de 1987 a été la première où on a décidé de faire une analyse de données longitudinales; cette analyse allait porter sur la période 1982-1987. Par la même occasion, on a décidé que les données de l'enquête allaient désormais être mises à la disposition du Soil Conservation Service de chaque État pour qu'il puisse faire ses propres analyses.

En 1987, les membres du personnel sur le terrain se sont vu remettre une feuille de travail qui contenait les données des segments pour 1982. Ils devaient y inscrire les données pour 1987 en se fondant sur les résultats d'observations sur le terrain et de la photographie aérienne. Ils étaient autorisés à corriger les données de 1982 si celles-ci étaient inexactes. Des méthodes de contrôle et de vérification ont été appliquées durant la phase de traitement.

On a conçu l'échantillon de manière à obtenir des estimations acceptables pour des unités appelées "zones principales de sols" (Major Land Resource Areas -- MLRA). Ces zones sont définies en fonction de la nature du sol et du couvert végétal. On en compte environ 180 sur le territoire visé par l'enquête. Par ailleurs, la superficie estimée pour chaque comté doit concorder avec la superficie totale du comté. L'échantillon de l'enquête comprend environ 3 100 comtés. Comme il doit y avoir concordance entre les estimations de superficie relatives aux comtés et celles relatives aux zones principales de sols, l'unité de totalisation fondamentale est la portion d'une zone principale de sols comprise dans les limites d'un comté. Ces unités de base sont au nombre de 5 530 et sont désignées par le sigle MLRAC.

Le plan de sondage équivaut à la forme la plus élémentaire d'une enquête par panel puisque l'échantillon de 1987 est à peu de choses près un sous-ensemble de l'échantillon de 1982. On a choisi d'utiliser comme variables de contrôle les superficies représentées en 1982 par 14 modes d'utilisation du sol parmi les plus importants (par ex. : terrain labourable, terrain de parcours, terrain forestier et terrain urbain). De plus, les données externes, comme la superficie représentée par les routes en 1987, et les données des segments, comme la superficie du terrain urbain en 1987, constituent de l'information supplémentaire au même titre que les données tirées des enregistrements incomplets.

Le tableau 4 est la version condensée d'un tableau d'estimations pour un des États visés par l'enquête. On n'y retrouve que 4 des 14 modes d'utilisation du sol considérés pour l'estimation. Les chiffres figurant dans la colonne de droite sont les estimations pour 1982. Les totaux des colonnes 3 (terrain urbain) et 4 (routes) sont tirés respectivement des données des segments et des sources externes. Le vecteur formé des quatre premiers totaux de la colonne de droite et des deux derniers totaux de la ligne du bas (superficie de terrain urbain en 1987 et superficie représentée par les routes en 1987) est un vecteur de totaux qui correspond au vecteur des moyennes estimées (x) de la section 3.

correspondants et c) l'échantillon de personnes qui ont participé aux quatre passages en 1985 de même que les poids correspondants.

Nous allons maintenant décrire une méthode d'estimation pour une enquête par panel où il y a des cas de non-réponse et où l'analyse n'est réalisée qu'à la toute fin. Nous supposons qu'une proportion raisonnable des unités participent à tous les passages de l'enquête et que l'analyse longitudinale revêt de l'intérêt. Il s'agit ici de construire des poids pour les unités dont les enregistrements sont complets. Les données fournies par les répondants dont les enregistrements sont incomplets servent d'information supplémentaire.

La première étape de l'analyse consiste à choisir quelques variables qui ont une grande importance pour l'enquête. Le nombre de variables que l'on peut utiliser dépendra de la taille de l'échantillon. Dans un deuxième temps, on calcule la structure des covariances du vecteur des estimations, qui est composé des estimations simples calculées pour chacune des variables pour chaque genre de schéma de réponse et chaque période pertinente. La structure des covariances est une fonction du schéma de réponse et de non-réponse. Il existe plusieurs définitions des estimateurs simples. Dans le cas de l'échantillonnage aléatoire simple, les estimateurs simples sont des moyennes. Dans le cas de l'échantilonnage stratifié, le vecteur initial peut être défini de manière à inclure les estimations pour chaque strate. Par ailleurs, l'estimateur simple pourrait servir à pondérer les réponses dans chaque strate pour compenser la non-réponse. Le vecteur Y figurant dans l'équation (1) est justement un vecteur d'estimations simples.

Étant donné le vecteur d'estimations simples et la matrice des covariances estimée de ce vecteur, nous pouvons construire à l'aide des moindres carres généralisés de meilleurs estimateurs pour chacune des périodes. Par exemple, si nous avons une enquête par panel avec trois passages, il y a sept schémas de réponse possibles, soit XXX, OXX, XOX, XXO, XOO, OXO et OOX, où X signifie réponse et O, non-réponse. Si nous choisissons deux variables d'intérêt, le vecteur des estimations simples contiendra 12 x 2 = 24 estimations puisqu'il y a 12 réponses - groupes associés aux sept schémas de réponse. Dans cet exemple, les moindres carrés généralisés serviraient à produire six estimations, soit les valeurs estimées des deux variables d'intérêt pour chacune des trois périodes.

L'estimateur par les moindres carrés généralisés des caractères choisis sert de variable de contrôle à une étape ultérieure. A l'aide de méthodes de régression, nous construisons des poids pour les personnes qui ont participé à tous les passages. Ces poids sont construits de telle manière que les estimations par les moindres carrés généralisés pour chaque période visée soient reproduites par l'échantillon pondéré des personnes qui ont participé aux trois passages. En d'autres termes, les valeurs estimées des variables choisies pour les diverses périodes servent de données de contrôle.

L'efficacité de cette méthode dépend de la corrélation entre les variables de contrôle choisies et la variable d'analyse. Si l'une des variables de contrôle est aussi la variable d'analyse, la méthode sera très efficace. La seule raison pour laquelle cette méthode n'est pas parfaitement efficace est que la méthode des moindres carrés généralisés n'utilise qu'une quantité limitée de données.

Son principal avantage est qu'elle produit une seule série de données de totalisation à partir de laquelle on peut construire des estimateurs ayant la propriété d'additivité pour toutes les périodes visées par l'enquête et tous les tableaux de variation brute.

La variance de cette méthode peut être calculée de la même manière que celle de la méthode utilisée pour l'échantillonnage double. Soit Y le caractère d'intérêt. Pour plus de simplicité, nous allons supposer un échantillonnage aléatoire simple à chaque fois. Nous définissons comme suit le modèle servant à l'estimation :

$$Y_i = \mu_Y + (X_i - \mu_X)\theta + e_i \ ,$$

$$\mu_X = E\{X\} \ ,$$

$$e_i \sim \mathrm{Ind}(0, \sigma_e^2) \ .$$

Soit $\hat{\mu}_X$ l'estimateur par les moindres carrés généralisés de μ_X. Alors, l'estimateur de la moyenne de Y s'écrit

$$\hat{\mu}_Y = \bar{y} + (\hat{\mu}_X - \bar{X})\hat{\theta} \ ,$$

où $(\bar{y}, \bar{x})$ est le vecteur de moyennes pour les éléments observés à tous les passages et θ est le vecteur des coefficients de régression que nous avons calculés en faisant la régression de Y(i) par rapport à X(i) à l'aide de la série d'enregistrements complets. Posons m comme le nombre de ces enregistrements. Alors, la variance de l'estimateur est approximativement

$$V\{\hat{\mu}_Y\} = m^{-1}\sigma_e{}^2 + \theta' V\{\hat{\mu}_X\}\theta \ ,$$

où $V\{\hat{\mu}_X\}$ est la matrice des covariances de $\hat{\mu}_X$.

Si nous remplaçons g par la combinaison linéaire GY , l'équation ci-dessus devient

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} Y \ .$$

Cette équation définit l'estimateur restreint de β comme une fonction linéaire de Y. Par conséquent, la variance de l'estimateur de β correspondra à la partie supérieure k x k de

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} V \left[\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} \right]'$$

Cette méthode n'est pas la seule façon de calculer l'estimateur par les moindres carrés généralisés restreint. Il y a aussi l'estimateur composite, qui est un autre estimateur de niveau et de variation qui ne modifie en rien l'estimateur précédent. Voir par exemple Wolter (1979).

Ce court exemple met en lumière plusieurs points. Premièrement, compte tenu d'une corrélation de 0.591 entre les niveaux d'emploi des deux périodes, l'application des moindres carrés généralisés améliore modérément (environ 10 %) l'estimation du niveau de chômage pour la période courante. En revanche, la même méthode améliore sensiblement la variance de la valeur estimée de P(EE). En effet, cette variance équivaut à environ 45 % de la variance de l'estimateur simple. Le second point à retenir est que l'application des moindres carrés généralisés avec contrainte donne des estimations de P(EE) et de P(.E) qui sont presque aussi efficientes que celles obtenues par les moindres carrés généralisés sans contrainte. Dans le cas de P(.E), la perte d'efficience est d'environ 1 % et dans le cas de P(EE), elle est d'environ 6 %.

3. ESTIMATEURS LONGITUDINAUX

Nous avons défini plus haut l'enquête à échantillon constant comme une enquête où les mêmes éléments sont observés à chaque période de collecte des données. L'enquête à échantillon constant se prête bien à l'observation de certaines unités physiques, comme des parcelles de terrain. Par contre, en ce qui a trait à l'observation de populations humaines, l'enquête à échantillon constant n'est rien de plus qu'une vue de l'esprit. Dans la réalité, l'enquêteur perd toujours une partie du groupe de répondants entre deux passages d'une enquête. Lepkowski (1989) et Little et Su (1989) font une bonne analyse des méthodes de traitement de la non-réponse. Voir aussi à ce sujet Little et Rubin (1987), Kalton (1983) et Madow et coll. (1983).

Nous avons aussi défini l'enquête avec renouvellement de l'échantillon, où certains éléments de l'échantillon font place à de nouveaux éléments à chaque passage de l'enquête. Dans ce cas, nous pouvons dire qu'il existe une forme de planification de la non-réponse pour les éléments qui sont supprimés de l'échantillon. Il faut donc voir un lien entre l'estimation en situation de non-réponse et l'estimation dans les enquêtes avec renouvellement partiel de l'échantillon.

Comme il est difficilement concevable qu'un enquêteur obtienne des réponses de tous les membres de l'échantillon à chaque passage de l'enquête, il faut s'attendre à recourir à une méthode qui permettra de compenser la non-réponse (prévue ou non prévue). Il existe deux méthodes simples et courantes. Si l'intention première de l'enquêteur est de suivre l'évolution d'un groupe de personnes dans le temps, très souvent il ne considérera dans son échantillon que les personnes qui ont participé à tous les passages de l'enquête. Dans ces circonstances, il dispose d'une méthode de pondération par laquelle il peut redresser les données de l'enquête à l'aide des caractéristiques du groupe initial des répondants et/ou de données supplémentaires. On procède souvent de cette façon dans les enquêtes spéciales portant sur une population spécifique. Dans ce cas, les résultats ne sont publiés qu'une fois l'enquête terminée.

La seconde méthode consiste à établir des estimations pour chaque période à l'aide des données dont on dispose pour la période en question. Cette méthode est souvent utilisée pour les enquêtes périodiques; les résultats sont publiés à la fin de chaque enquête et ne sont pas révisés par la suite et aucune estimation longitudinale n'est produite. Un des avantages de cette méthode est qu'il est très facile d'établir des estimations pour la période t puisqu'on ne se sert pas des données de la période précédente pour calculer les estimations des valeurs courantes. Avec cette méthode, on obtient habituellement des estimations acceptables (non optimales) des valeurs courantes mais des estimations de la variation qui laissent à désirer.

Par ailleurs, on peut utiliser les deux méthodes dans une même enquête. La Survey of Income and Program Participation (SIPP), réalisée par le U.S. Bureau of the Census, est une enquête par panel qui prévoit un renouvellement partiel de l'échantillon à chaque période d'interview, c'est-à-dire à tous les quatre mois. À chaque passage de l'enquête, le U.S. Bureau of the Census produit une série de poids qui peuvent servir à établir des estimations pour la période en question à l'aide des données fournies par l'ensemble des personnes qui ont participé à l'enquête à cette occasion. L'organisme américain produit aussi a) l'échantillon de personnes qui ont participé aux huit passages de l'enquête en 1984-1985 de même que les poids relatifs à ces personnes, b) l'échantillon de personnes qui ont participé aux quatre passages de l'enquête en 1984 de même que les poids

Tableau 3. Variances obtenues avec diverses méthodes d'estimation (pour un échantillon de taille n à chaque période, multiplier chaque élément du tableau par 2, puis diviser par n.)

Paramètre	Méthode		
	Simple	MCG avec contrainte	MCG
$P_{E.}$	0.0326	0.0326	0.0294
P_{EE}	0.0819	0.0397	0.0374
P.E	0.0278	0.0258	0.0255
PEE/P.E	0.0290	0.0229	0.0220
P.E - PE.	0.0429	0.0367	0.0353

La dernière colonne du tableau 3 contient les variances du meilleur estimateur linéaire sans biais construit à l'aide des moindres carrés généralisés. Cet estimateur est construit a l'aide du vecteur des cinq paramètres statistiques fondamentaux et de la matrice des covariances de ce vecteur. Sa formulation est la suivante :

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \ , \tag{1}$$

où V est définie dans le tableau 2, $\beta = (P_{E.} \ , P_{.E}, P_{EE})$,

$$X' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

et Y est le vecteur quinquidimensionnel des estimations directes,

$$Y' = (\tilde{P}_{E.1}, \tilde{P}_{E.2}, \tilde{P}_{EE}, \tilde{p}_{.E2}, \tilde{P}_{.E3}) \ .$$

La seconde colonne du tableau 3 contient les variances de l'estimateur par les moindres carrés restreint, la contrainte étant que l'estimateur pour la période 1 doit être l'estimateur obtenu à l'aide de l'échantillon initial. Cette condition serait appropriée si un organisme statistique ne révisaient jamais les estimations déjà publiées. Par exemple, le Bureau of Labour Statistics des États-Unis ne révise jamais les statistiques du chômage. Une fois publiées, ces statistiques tiennent lieu d'estimations officielles. Il faut préciser toutefois qu'elles reposent sur un échantillon plus complexe et une enquête qui s'étend sur une plus longue période.

Pour décrire l'estimateur par les moindres carrés généralisés restreint du tableau 3, définissons le modèle

$$Y = X\beta + e \ ,$$

où X est une matrice fixe n x k et

$$E(ee') = V \ .$$

L'estimateur par les moindres carrés généralisés de β , dont certains éléments sont contraints à être des combinaisons linéaires de Y , peut être construit de la façon suivante. Considérons la fonction lagrangienne

$$(Y - X\beta)' \ V^{-1}(Y - X\beta) - 2 \sum_{i=1}^{b} \lambda_i(r_i\beta - g_i)$$

où r(i) est un vecteur ligne fixe et b est le nombre de contraintes. La solution à ce problème de minimisation est définie

$$\begin{pmatrix} X'V^{-1}X & r' \\ r & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X'V^{-1}Y \\ g \end{pmatrix},$$

où $\lambda' = (\lambda_1,\lambda_2,\ldots,\lambda_b)$, $r' = (r_1',r_2',\ldots,r_b')$ et $g' = (g_1, g_2, \ldots, g_b)$.

l'organisme qui publie les données peut être tenu de respecter un plafond en ce qui concerne le nombre de fois où il est permis de réviser des estimations antérieures. Smith et Holt (1989) se sont penchés sur ce dernier point.

Afin d'illustrer ces problèmes d'estimation, nous avons voulu utiliser un exemple simple. À cette fin, le tableau 1 représente un tableau de contingence qui montre la division de la même variable en deux classes pour deux périodes données, et dont les observations reposent sur un très grand échantillon. Nous avons identifié les classes de ce tableau en désignant la première comme les personnes avec emploi et la seconde comme les chômeurs. Nous supposons que la population ne varie pas d'une année à l'autre. Si nous devions considérer les naissances et les décès, il nous faudrait alors un tableau 3 x 3. Supposons que nous voulons estimer la variation de niveau d'une période à l'autre. Supposons aussi que nous voulons dresser un tableau des variations brutes; cette opération suppose l'estimation des fréquences par case du tableau de contingence. Pour un tableau 2 x 2, il suffit d'estimer la fréquence de la case (1,1) et les proportions marginales pour obtenir les fréquences des autres cases.

Tableau 1. Proportions hypothétiques pour deux périodes données

PÉRIODE 1	PÉRIODE 2		
	Personnes avec emploi	Chômeurs	Total
Personnes avec emploi	0.91	0.02	0.93
Chômeurs	0.03	0.04	0.07
Total	0.94	0.06	1.00

Notre analyse porte sur deux périodes, pour lesquelles le même nombre d'éléments sont observés. Nous supposons que la moitié des éléments observés à la première période le sont aussi à la seconde. Autrement dit, les éléments observés à la seconde période se répartissent en deux groupes égaux : un groupe formé d'éléments observés à la première période et l'autre formé de nouveaux éléments. Le vecteur des observations est composé de la proportion d'éléments de la classe 1 qui font partie de la moitié de l'échantillon observée uniquement à la première période [désignée par P(E.1)], de la proportion d'éléments de la classe 1 qui font partie de l'autre moitié de l'échantillon de la période 1 [désignée par P(E.2)], de la proportion d'éléments de la classe 1 aux deux périodes, qui font partie de la moitié d'échantillon observée aux deux périodes [désignée par P(EE)], de la proportion d'éléments de la classe 1 à la période 2 parmi les éléments observés aux deux périodes [désignée par P(.E2)] et de la proportion d'éléments de la classe 1 à la période 2 parmi les éléments qui n'ont été observés qu'à la période 2 [désignée par P(.E3)]. Les arguments figureront entre parenthèses dans le texte et figureront comme indices inférieurs dans les tableaux.

Nous supposons un échantillonnage aléatoire simple. Comme les statistiques consistent en des proportions d'échantillon, on peut reproduire facilement la matrice des covariances du vecteur formé de cinq estimations. Un multiple de cette matrice est représenté par le tableau 2. Pour obtenir la matrice des covariances pour un échantillon de taille n à chaque période, il suffit de diviser chaque élément de la matrice du tableau 2 par n, puis de multiplier le résultat par deux. Le tableau 3 donne les variances obtenues avec diverses méthodes d'estimation. La première colonne contient les variances obtenues avec la méthode qui n'utilise que les éléments de l'échantillon de la première période pour estimer la proportion d'éléments de la classe 1 (personnes avec emploi) à cette période. Pour estimer la proportion d'éléments de la classe 1 aux deux périodes, la méthode ordinaire n'utilise que les éléments communs aux deux échantillons et pour estimer la proportion d'éléments de la classe 1 à la période t, elle n'utilise que l'échantillon observé à la période 2. Par conséquent, si nous avons un échantillon de 200 éléments à chaque période, l'échantillon de la première période sert à estimer la proportion d'éléments de la classe 1 à cette période, les 100 éléments communs aux deux périodes servent à estimer la proportion d'éléments qui demeurent dans la classe 1 d'une période à l'autre et les 200 éléments observés à la période 2 servent à estimer la proportion d'éléments de la classe 1 pour cette période.

Tableau 2. Matrice des covariances du vecteur des proportions d'échantillon; deux périodes et échantillons se chevauchant dans une proportion de 50 %. (Pour un échantillon de taille n, multipler chaque élément du tableau par 2, puis diviser par n.)

$P_{E.1}$	$P_{E.2}$	P_{EE}	$P_{.E2}$	$P_{.E3}$
0.0651	0	0	0	0
0	0.0651	0.0637	0.0358	0
0	0.0637	0.0819	0.0546	0
0	0.0358	0.0546	0.0564	0
0	0	0	0	0.0564

d'une fois, 2) l'enquête à échantillon constant, où les mêmes éléments sont observés à chaque période, 3) l'enquête avec renouvellement de l'échantillon, où des éléments de la population sont observés pour un nombre déterminé de périodes, puis supprimés de l'échantillon par renouvellement selon un plan déterminé, et 4) l'enquête à panel fractionné, qui est une combinaison de l'enquête à échantillon constant et de l'enquête 1) ou 3). Duncan et Kalton indiquent aussi sous forme de tableau les genres d'enquêtes qui conviennent le mieux aux différents objectifs.

Lorsqu'un établissement réalise une enquête à passages répétés, il doit parer à toutes les difficultés qui accompagnent normalement l'exécution d'enquêtes sauf que dans ce cas-ci, les problèmes sont amplifiés. La non-réponse demeure un sujet de préoccupation mais il est plus difficile d'obtenir la collaboration constante des répondants pour plusieurs périodes successives. L'erreur de réponse est aussi présente dans ce genre d'enquête sauf qu'il faut composer avec un phénomène de "conditionnement" lié à la répétition des interviews. De plus, les erreurs de réponse ont pour effet de créer des incohérences dans les données lorsque celles-ci sont recueillies sur une longue période. Pour assurer la qualité d'une enquête à passages répétés, il est nécessaire de procéder toujours de la même façon sur le terrain et d'appliquer les mêmes méthodes de traitement et d'estimation pour toutes les périodes. Par ailleurs, la gestion des données soulève plus de difficultés lorsqu'il s'agit d'enquêtes à passages répétés. Enfin, le changement de composition des unités, telles les familles, vient compliquer l'estimation et l'analyse.

Nous n'aborderons ici que quelques-unes des questions qui se rattachent aux enquêtes à passages répétés. Notre analyse est fondée sur une grande enquête réalisée par le U.S. Soil Conservation Service en collaboration avec l'université Iowa State. Dans la section 2, nous examinons quelques-unes des méthodes d'estimation utilisées dans les enquêtes à passages répétés. Cette analyse se prolonge dans la section 3, où il est surtout question de l'estimation de paramètres longitudinaux dans les enquêtes par panel. Dans la section 4, nous exposons brièvement les méthodes d'estimation utilisées dans l'enquête du U.S. Soil Conservation Service. Enfin, la section 5 renferme une brève description des effets de l'erreur de mesure sur les estimations de la variation brute.

2. ESTIMATION

Dans cette section, nous allons exposer à grands traits la méthode d'estimation par les moindres carrés généralisés appliquée à des enquêtes où seul un sous-ensemble des éléments de l'échantillon est observé pendant des périodes consécutives. La méthode des moindres carrés généralisés est la première méthode à laquelle se sont intéressés les auteurs qui étudiaient l'estimation dans les enquêtes à passages répétés. Sur les traces de Cochran (1942), Jessen (1942) fut le premier à envisager la construction de poids à variance minimum pour une série d'estimateurs non biaisés établis pour chaque période visée par l'enquête.

Jessen (1942) a analysé le cas particulier de l'échantillonnage effectué à deux reprises où le nombre d'observations diffère d'un échantillon à l'autre et s'est intéressé à la répartition optimale des unités entre les groupes d'échantillons chevauchants et non chevauchants. Patterson (1950) a examiné le cas de T sondages successifs avec plusieurs modes de renouvellement partiel des unités. Le plan d'échantillonnage le plus simple prévoyait le renouvellement d'une proportion déterminée des unités d'échantillonnage à chaque nouveau sondage. En outre, Patterson (1950) avait supposé que, pour une unité i donnée, les écarts $x(ti) - x(t)$, $t = 1, 2, \ldots$, suivaient un processus autorégressif du premier ordre, $x(ti)$ étant la valeur de l'unité de population i au temps t et $x(t)$, la moyenne de la population finie correspondante. Suivant le modèle d'erreur qui en a découlé, il a défini des estimateurs optimaux des valeurs fixes $x(t)$ et des écarts $x(t) - x\ (t-1)$. Il s'est également penché sur l'estimation optimale de $x(t)$ suivant des formes généralisées du plan de renouvellement partiel, la détermination de la taille optimale de l'échantillon et l'estimation avec erreurs non autorégressives.

La méthode des moindres carrés a été approfondie par Eckler (1955), Gurney et Daly (1965) et Jones (1980). On en est venu aussi à parler d'estimateurs composites; voir à ce sujet Rao et Graham (1964), Graham (1973) et Wolter (1979). Battese, Hasabelnaby et Fuller (1989) décrivent comment le Département de l'agriculture des É.-U. applique la méthode des moindres carrés dans son enquête sur les activités des exploitations agricoles.

Il semble juste d'affirmer que ces auteurs se sont intéressés surtout à des moyennes ou à des totaux pour des périodes précises. Autrement dit, ils n'ont pas étudié explicitement des paramètres longitudinaux comme la proportion d'individus appartenant à une classe particulière à la période 1 et à la période 2. Nous verrons toutefois que la méthode des moindres carrés s'applique à des paramètres de ce genre.

Une caractéristique intéressante de la méthode des moindres carrés linéaires est que les estimateurs relatifs à un certain nombre de caractères ont la propriété d'additivité, c'est-à-dire que la somme de l'estimateur par les moindres carrés de $\bar{Y}$ et de l'estimateur par les moindres carrés de $\bar{Z}$ est égale à l'estimateur par les moindres carrés de $\bar{Y} + \bar{Z}$. Toutefois, si l'on se sert d'autres vecteurs d'observations pour construire des estimateurs, la propriété d'aditivité disparaît.

Dans beaucoup d'enquêtes, on ne peut calculer les estimateurs par moindres carrés optimaux pour toutes les périodes. D'abord, on ne peut se servir de toute l'information disponible pour l'estimation, c'est-à-dire qu'on ne peut intégrer toutes les données des enquêtes des périodes antérieures à une analyse par les moindres carrés pour la période courante. Souvent, le nombre de variables dépassera le nombre d'observations. Ensuite,

Recueil du symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

ANALYSE D'ENQUÊTES À PASSAGES RÉPÉTÉS

W.A. Fuller[1]

RÉSUMÉ

Dans cet article, nous nous intéressons principalement aux enquêtes à passages répétés où une partie des unités de l'échantillon est observée sur plusieurs périodes et une partie n'est pas observée à certaines périodes. Nous voyons en quoi consiste l'estimation par les moindres carrés pour de telles enquêtes. Nous nous arrêtons aussi à des méthodes d'estimation, modifiées de telle manière que les estimations existantes n'ont pas à être révisées lorsque de nouvelles données sont connues. Par ailleurs, nous considérons des méthodes pour estimer des paramètres longitudinaux; mentionnons à cet égard les tableaux de variation brute. Nous décrivons aussi la méthode d'estimation utilisée dans une enquête à passages répétés sur l'utilisation du sol, réalisée par le U.S. Soil Conservation Service. Enfin, nous illustrons l'effet de l'erreur de mesure sur les estimations de la variation brute et montrons qu'un plan de sondage qui permet d'estimer les paramètres du processus d'erreur de mesure peut être très efficient.

MOTS CLÉS : Échantillon d'enquête, moindres carrés, erreur de mesure, variation brute.

1. INTRODUCTION

L'analyse d'enquêtes à passages répétés suscite beaucoup d'intérêt. Soulignons à cet égard la publication récente des actes d'un symposium sur les enquêtes par panel, colligés par Kasprzyk, Duncan, Kalton et Singh (1989), la tenue de séances sur la question lors des deux dernières assemblées de l'Institut international de Statistique, et le présent symposium. Dansl'article qu'ils ont présenté à la session de l'IIS de 1989 à Paris, Smith et Holt (1989) parlent d'un intérêt renouvelé pour l'élaboration et l'analyse d'études longitudinales. Ils soulignent que des spécialistes de domaines comme la sociologie et la santé réalisent depuis longtemps des enquêtes par panel et des études de cohorte. Ils citent Lazarsfeld et Fiske (1938). Dans le domaine de la santé, mentionnons l'article de Garcia, Battese et Brewer (1975).

Les organismes officiels réalisent de nombreuses enquêtes périodiques, comme l'enquête sur la population active. Ces enquêtes produisent habituellement une suite de rapports comme ceux portant sur l'emploi et le chômage pour la période courante. En règle générale, les enquêtes réalisées par les organismes officiels fournissent très peu de données sur le comportement des unités de l'échantillon dans le temps. La U.S. Survey of Income and Program Participation est un exemple d'enquêtes qui servent à produire des estimations longitudinales. Voir à ce sujet Kasprzyk et McMillen (1987). Bien que nous en sachions moins sur les enquêtes réalisées par le secteur privé que sur celles réalisées par les administrations publiques, il semble que les premières servent surtout, comme les secondes, à produire une suite de rapports pour des périodes données. Toutefois, le secteur public comme le secteur privé doivent répondre à une demande accrue d'analyses longitudinales.

L'élaboration d'une taxinomie pour les enquêtes à passages répétés a pour effet de mettre en relief les questions complexes qui accompagnent ce genre d'enquêtes. Duncan et Kalton (1987) énumèrent sept objectifs des enquêtes à passages répétés, soit :

A. produire des estimations de paramètres de la population pour des périodes déterminées;
B. produire des estimations de paramètres de la population pour des périodes combinées;
C. mesurer la variation nette;
D. mesurer des éléments de la variation, dont
   i) la variation brute
   ii) la variation pour une unité
   iii) la variabilité pour une unité
E. produire des données agrégees sur les unités prises individuellement
F. déterminer la fréquence, le moment et la durée d'événements
G. accumuler des données sur des populations peu courantes.

Bien que cela ne soit pas explicite, plusieurs de ces objectifs supposent l'estimation des paramètres de modèles spécialisés.

Par ailleurs, Duncan et Kalton définissent quatre genres d'enquêtes : 1) l'enquête à passages répétés, où rien n'est fait pour veiller à ce que des éléments particuliers de la population fassent partie de l'échantillon plus

[1] Department of Statistics, Iowa State University, Ames (Iowa).

SECTION 1

L'ÉCHANTILLONNAGE RÉPÉTÉ

Pour qu'elle puisse prospérer, la communauté des théoriciens doit aborder des problèmes réels et importants. Les praticiens peuvent les fournir. La plus grande partie du travail d'un praticien porte sur le découpage et l'application de la théorie à des fins précises, ce qui permettra de mettre en évidence les limites éventuelles de la théorie existante, posant d'autres défis au théoricien.

Je dois dire quelques mots ici sur la raison de l'importance de ce sujet à Statistique Canada et pour les autres organismes statistiques. Presque toutes les données que nous publions sont des séries temporelles. Il n'existe pas beaucoup de statistiques dont on pourrait dire que le seul intérêt qu'elles présentent est leur valeur aujourd'hui. Les gens veulent savoir pourquoi les choses évoluent, et cela donne les séries temporelles, que nous les appelions ainsi ou pas. Quelles sont donc les tendances générales qui font que le thème de ce symposium est particulièrement important?

Comme tout le monde, nous faisons face à des contraintes en matière de ressources. Nous voulons par conséquent extraire le maximum d'information des données existantes sans une collecte supplémentaire onéreuse de données. La dimension temps dans l'analyse peut aider.

En tant que notre principale source de données, les enquêtes doivent être concices de façon optimals. On peut obtenir des avantages appréciables en tenant compte de la dimension temporelle, lors de la conception et de l'estimation, pour les enquêtes dont les données serviront à suivre l'évolution d'un phénomène.

Un autre sujet de préoccupations est la compréhension et l'interprétation par les utilisateurs des données que nous publions. Certains des aspects les moins bien compris de nos données se rapportent au temps. Je fais référence ici à la désaisonnalisation et aux révisions qui incorporent des données plus tardives aux séries publiées plus tôt sous une forme provisoire. Nous estimons qu'il y a des progrès à faire, sinon sur le plan de la simplification des procédures, au moins sur celui des explications et de la garantie de l'obtention d'ensembles de données cohérentes.

Enfin, il y a l'intérêt croissant que l'on observe à l'égard des données longitudinales au micro-niveau, c'est-à-dire des renseignements sur l'évolution des unités (personnnes, entreprises, fermes, etc.) plutôt que sur simplement l'évolution des agrégats. Là encore les méthodes des séries temporelles peuvent se révéler utiles.

Telles sont quelques-unes des questions auxquelles nous devons répondre aujourd'hui et qui rendent le thème de ce Symposium important pour nous.

Le programme semble très intéressant, avec une bonne combinaison de théorie et de pratique dans divers domaines dont la démographie, l'économétrie, l'éducation et l'épidémiologie. J'espère que vous allez tous bénéficier de ce Symposium et que certains d'entre vous seront inspirés pour poursuivre davantage l'élaboration ou l'application de la théorie dans ce domaine. J'espère également que l'extension de la collaboration entre les statisticiens universitaires et publics suscitera de l'intérêt.

Je vous remercie de votre participation à ce Symposium et je vous souhaite trois journées intéressantes et productives.

Recueil du Symposium de Statistique Canada
sur l'analyse des données dans le temps
octobre 1989

INTRODUCTION

G.J. Brackstone[1]

Au nom de Statistique Canada, je vous souhaite la bienvenue à Symposium 89. Ce Symposium est organisé conjointement par Statistique Canada, le laboratoire de recherche en statistique et probabilités de l'Université Carleton et l'Université d'Ottawa. Il est très réconfortant de voir autant de monde ce matin. Ceci prouve que nous avons choisi un sujet très d'actualité et élaboré un programme intéressant, ou encore, que notre comité organisateur a particulièrement bien réussi son effort de commercialisation, ou les deux.

Le thème de ce Symposium est l'analyse des données dans le temps. Ce titre, naturellement, est quelque peu ambigu, du moins en anglais. Ceux d'entre vous qui sont venus ici pour apprendre comment accélérer leur analyse ou garantir le respect des délais, risquent d'être décus, parce que ce n'est pas dans ce sens-là que nous avons interprété le thème de ce Symposium. Le thème principal cette année est la collecte, le traitement et, en particulier, l'analyse des données dans le temps.

Toujours dans cet ordre d'idées, permettez-moi de préciser que ce Symposium est le sixième d'une série de symposiums sur la méthodologie à Statistique Canada. Le sujet des symposiums précédents de 1984 jusqu'en 1988 ont été "L'Analyse des données d'enquêtes" (1984), où nous nous sommes intéressés à l'analyse transversale des données provenant d'enquêtes complexes; "Les statistiques régionales" (1985), ce qui a permis la publication d'un livre; Les données manquantes dans les enquêtes (1986), symposium de moindre envergure, mais avec quelques conférenciers de haut calibre qui ont traité de ce problème pour les organismes statistiques; Les utilisations statistiques des données administratives (1987), lorsque nous avons accueilli un ensemble vraiment international de conférenciers qui ont parlé des aspects statistiques et des problèmes d'intrusion dans la vie privée due à l'utilisation des données administratives; et enfin l'an dernier, Les repercussions de la technologie de pointe sur les enquêtes, qui nous a donné l'occasion d'étudier la synergie entre la méthodologie d'enquêtes et l'informatique.

Je laisserai aux enthousiastes le soin de déterminer si la série temporelle de ces symposiums constitue un événement aléatoire ou non aléatoire. Pour les spécialistes de la prédiction, je leur laisse le défi de prédire le sujet du symposium de l'année prochaine avant qu'il ne soit annoncé plus tard au cours de la semaine.

Le sujet pour cette année, L'analyse des données dans le temps, représente, à mon avis, un choix très opportun et à propos. Il nous fournit l'occasion d'écouter un échange d'idées entre les théoriciens et les praticiens, et aussi entre les statisticiens des universités et ceux ou celles des gouvernements et d'autres agences. En dépit des développements importants en ce qui concerne la théorie ou la pratique des méthodes de séries chronologiques, et en dépit de la disponibilité des données qui découlent des expériences répétées, des enquêtes régulières, des recensements, et des fichiers administratifs, il existe les méthodes de séries chronologiques avec les attributs bien-connus et avantageux, qu'on n'utilise presque jamais, et certainement pas d'une façon routinière, dans les programmes des agences gouvernementales.

Il y a peut-être trois causes principales à cette situation:

D'abord, ces méthodes comportent souvent des calculs et des manipulations de données assez complexes, et un fardeau de computation assez lourds;

Deuxièmement, il y a peut-être un manque de connaissance, chez les praticiens, en particulier ceux qui s'occupent des enquêtes, de la théorie qui existe actuellement;

Troisièmement, il y a, bien sur, des lacunes et des faiblesses dans la théorie - ele ne répond pas à toutes les situations auxquelles les praticiens font face.

La première de ces causes, la complexité des calculs, même si elle ne doit pas être laissé entièrement de côté, n'est pas le thème dominant aujourd'hui, et elle perdra probablement encore davantage de son importance dans l'avenir. Mais les deux autres causes, à savoir la méconnaissance de la théorie chez les praticiens et les lacunes de la théorie, persisteront jusqu'à ce que nous trouvions une solution. Et c'est la raison d'être de ce symposium. C'est l'une des façons qui nous permettra de combler l'écart entre la théorie et la pratique, entre les théoriciens et les praticiens.

[1] G.J. Brackstone, Secteur de l'informatique et de la méthodologie, Statistique Canada, Ottawa, (Ontario) K1A OT6

ALLOCUTION D'OUVERTURE

D. KREWSKI, Centre d'hygiène du milieu, Santé et Bien-être Social Canada, Ottawa, Ontario, Canada.

N. LANIEL, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, Canada.

J.F. LAWLESS, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

P. LIN, Division des mesures et analyse des industries, Statistique Canada, Ottawa, Ontario, Canada.

L. LIU, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

I.B. MacNEILL, Department of Statistical and Actuarial Sciences and Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada.

R.H. McGUCKIN, Center for Economic Studies, U.S. Bureau of the Census, Washington, D.C., U.S.A.

B.C. MONSELL, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

C. NAIR, Centre canadien d'information sur la santé, Statistique Canada, Ottawa, Ontario, Canada.

M. NARGUNDKAR, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada.

T. PETERSEN, Division des mesures et analyse des industries, Statistique Canada, Ottawa, Ontario, Canada.

D. PFEFFERMANN, Hebrew University, Jerusalem, Israel.

R. PRESSAT, Département de la Conjoncture, Institut national d'études démographiques, Paris, France.

M.G. PUGH, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.

J.N.K. RAO, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

G.R. ROBERTS, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada.

D. ROBERTSON, Division des études sociales et économiques, Statistique Canada, Ottawa, Ontario, Canada.

G. ROWE, Division des études sociales et économiques, Statistique Canada, Ottawa, Ontario, Canada.

L. SAGER, Division des mesures et analyse des industries, Statistique Canada, Ottawa, Ontario, Canada.

J. SHEDDEN, Centre d'hygiène du milieu, Santé et Bien-être Social Canada, Ottawa, Ontario, Canada.

A.C. SINGH, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada.

J. STRACHAN, School of Health Information Science, University of Victoria, Victoria, Columbie-Britannique, Canada.

T. WANNELL, Groupe d'analyse des entreprises et du marché du travail, Direction des études analytiques, Statistique Canada, Ottawa, Ontario, Canada.

M.C. WOLFSON, Division des études sociales et économiques, Statistique Canada, Ottawa, Ontario, Canada.

K.M. WOLTER, A.C. Nielsen, Northbrook, Illinois, U.S.A.

AUTEURS

J.R. BALDWIN, Groupe de l'analyse des entreprises et du marché du travail, Statistique Canada, Ottawa, Ontario, Canada, et professeur d'économie, Queen's University, Kingston, Ontario, Canada.

W.R. BELL, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

D.R. BELLHOUSE, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

S. BEN-TUVIA, Central Bureau of Statistics, Jerusalem, Israel.

D.A. BINDER, Division des méthodes d'enquêtes - entreprises, Statistique Canada, Ottawa, Ontario, Canada.

G.J. BRACKSTONE, Secteur de l'informatique et de la méthodologie, Statistique Canada, Ottawa, Ontario, Canada.

D.R. BRILINGER, Department of Statistics, University of California, Berkeley, California, U.S.A.

L. BUREK, Central Bureau of Statistics, Jerusalem, Israel.

R.T. BURNETT, Centre d'hygiène du milieu, Santé et Bien-être Social Canada, Ottawa, Ontario, Canada.

P.A. CHOLETTE, Division de la recherche et de l'analyse de séries chronologiques, Statistique Canada, Ottawa, Ontario, Canada.

G.H. CHOUDHRY, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, Canada.

W. CLARK, Division de l'éducation, de la culture et du tourisme, Statistique Canada, Ottawa, Ontario, Canada.

J.P. DICK, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada.

Q.P. DUONG, Bureau of Management Consulting, Ottawa, Ontario, Canada.

D.F. FINDLEY, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., U.S.A.

C. FORTIER, Division de la démographie, Statistique Canada, Ottawa, Ontario, Canada.

W.A. FULLER, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.

K. FYFE, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, Canada.

J.F. GENTLEMAN, Division des études sociales et économiques, Statistique Canada, Ottawa, Ontario, Canada.

P.K. GORECKI, Conseil économique du Canada, Ottawa, Ontario, Canada.

R.M. HARTER, A.C. Nielsen, Northbrook, Illinois, U.S.A.

V.K. JANDHYALA, Department of Mathematics, Washington State University, Pullman, Washington, U.S.A.

H. JOHANSEN, Direction de la promotion de la santé, Santé et Bien-être Social Canada, Ottawa, Ontario, Canada.

J.D. KALBFLEISCH, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

L. KISH, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, U.S.A.

SECTION 5. ÉPIDÉMIOLOGIE

SECTION 6. DÉMOGRAPHIE

SECTION 7. ÉCONOMÉTRIE

SECTION 8. ÉDUCATION

ALLOCUTION DE CLÔTURE

TABLE DES MATIÈRES

ALLOCUTION D'OUVERTURE

SECTION 1. L'ÉCHANTILLONNAGE RÉPÉTÉ

SECTION 2. L'ANALYSE DES SÉRIES CHRONOLOGIQUES EN PRÉSENCE D'ERREURS D'ENQUÊTES

SECTION 3. L'ANALYSE DES SÉRIES CHRONOLOGIQUES DE COMPTES

SECTION 4. DÉVELOPPEMENTS DANS L'ANALYSE DE DONNÉES DES SÉRIES CHRONOLOGIQUES

Préface

Au cours des dernières années, il y a eu une demande croissante de la part des secteurs gouvernementaux et privés, pour des outils statistiques appropriés pour mener des analyses de données recueillies périodiquement dans le temps à partir d'enquêtes, de recensements ou de sources administratives. Face à cette demande, un symposium international sur l'Analyse des données dans le temps a été organisé afin de rassembler des chercheurs et des praticiens de plusieurs domaine provenant des universités, du gouvernement, et d'autres organismes statistiques. Le symposium a été organisé par Statistique Canada et le Laboratoire de recherche en statistique et probabilité, Carleton University et l'Université d'Ottawa.

Le symposium s'est déroulé à Ottawa, du 23 au 25 octobre 1989, à la salle de conférences Simon Goldberg, à Statistique Canada; 325 participants enregistrés y ont assisté. Plusieurs exposés de statisticiens bien connus à travers le monde ont été présentés. Le discours-programme a été donné par le professeur Wayne Fuller de Iowa State University. Le conférencier-invité spéciale, le professeur David Brillinger de University of California at Berkeley, n'a pas pu se présenter à cause des circonstances difficiles causées par le tremblement de terre en Californie, mais son exposé est cependant inclut dans le recueil pour le bénéfice des lecteurs.

Le présent volume contient 27 exposés avec un contenu varié de théorie et de pratique. La grande variété de sujets traités au symposium devrait être très utile pour des chercheurs et praticiens impliqués dans divers champs des statistiques. Les exposés sont regroupés selon les huit sections suivantes:

Section 1: L'échantillonnage répété
Section 2: L'analyse des séries chronologiques en présence d'erreurs d'enquêtes
Section 3: L'analyse des séries chronologiques de comptes
Section 4: Développements dans l'analyse de données des séries chronologiques
Section 5: Épidémiologie
Section 6: Démographie
Section 7: Économétrie
Section 8: Éducation

Le recueil contient aussi l'allocution d'ouverture faite par G. Brackstone, et l'allocution de clôture donné par D.Binder. La traduction française des exposés a été révisée par plusieurs méthodologistes. Nos remerciements vont à: J. Armstrong, S. Beaulieu, J.-M. Berthelot, J.-R. Boudreau, R. Boyer, M. Brodeur, M. Bureau, P. Daoust, P. David, J. Denis, J. Dufour, J. Dumais, S. Giroux, M. Joncas, M. Lachance, D. Lalande, E. Langlet, Y. Leblond, J. Lynch, C. Morin, S. Perron, C. Poirier, G.Sampson, P. St-Martin, A. Théberge, M. Thibeault, et J. Tourigny. C'est avec plaisir que nous remercions aussi Judy Clarke, Carole Jean-Marie, Christine Larabie, Carmen Lacroix et Pat.Pariseau pour l'efficacité avec laquelle le travail s'est effectué, et spécialement Judy pour la coordination du travail.

L'organisation du symposium a été rendue possible grâce aux efforts de plusieurs personnes de Statistique Canada, en particulier ceux de J. Mayda et J. Morabito. Nous désirons aussi remercier D. Binder, G. Brackstone, D. Drew, J. Kovar, J.N.K. Rao, et M.P. Singh pour leurs encouragements et consultations. Finalement, notre appréciation doit être offerte à tous les conférenciers qui ont fait un grand succès du symposium.

A.C. Singh
P. Whitridge
Comité organisateur et éditorial
Symposium '89

Ottawa, (Ontario), Canada
Octobre 1990

Prix: 20.00$ Canadien

Payable au
"Receveur général du Canada - Recueil du Symposium '89"

Envoyer votre demande à:

Statistique Canada Recueil du Symposium '89
Division des méthodes des enquêtes sociales
Édifice R.H. Coats
Parc Tunney
Ottawa, (Ontario)
Canada
K1A 0T6

L'ANALYSE DES DONNÉES DANS LE TEMPS

Recueil du symposium international de 1989

Édité par

A.C. Singh et P. Whitridge
Direction de la méthodologie
Statistique Canada
Ottawa, (Ontario), Canada

Statistique Canada Statistics Canada

Canada

LA SÉRIE DES SYMPOSIUMS DE MÉTHODOLOGIE

1984	L'analyse des données d'enquête
1985	Les statistiques sur les petites régions
1986	Les données manquantes dans les enquêtes
1987	Les utilisations statistiques des données administratives
1988	Les répercussions de la technologie de pointe sur les enquêtes
1989	L'analyse des données dans le temps
1990	Mesure et amélioration de la qualité des données (à venir)

L'ANALYSE DES DONNÉES DANS LE TEMPS

L'ANALYSE DES DONNÉES DANS LE TEMPS

RECUEIL DU SYMPOSIUM INTERNATIONAL DE 1989

Édité par

A.C. Singh et P. Whitridge

Statistique Canada
Statistics Canada

Canada