Catalogue 12-001

Survey Methodology

A Journal of Statistics Canada

December 1990 Volume 16 Number 2

Statistics Canada Statistique Canada

Canada

Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

December 1990 Volume 16 Number 2

Published under the authority of the Minister
of Industry, Science and Technology

March 1991

Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

MORRIS H. HANSEN
(1910-1990)

This issue is dedicated to the memory of Morris H. Hansen,
a pioneer, innovator and leader
who made fundamental and lasting contributions
to many aspects of survey methodology.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 16, Number 2, December 1990

CONTENTS

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 165–166
Statistics Canada

In This Issue

This issue contains a special section on time series methods in surveys, a topic that has attracted considerable interest in recent years. Special thanks are due to W.A. Fuller and J.N.K. Rao for coordinating the editorial work for this section.

The first two papers of the special section deal with the problems of sample design and maintenance, and estimation of various parameters of interest in repeated surveys. Fuller notes that repeated surveys designed to enable estimation of the parameters of the measurement error process can be very cost efficient. For a two-period survey with fifty percent overlap, he shows that generalized least square estimates of longitudinal parameters can have substantially lower variance than the simple estimator based only on the overlapping units. Wolter and Harter deal with the problem of sample maintenance for a recurring survey. The ingenious use of a Peano curve allows the sample maintenance to meet several desirable properties. They describe an application to a marketing survey.

Bell and Hillmer discuss the underlying philosophy of the time series approach to estimation in repeated surveys based on the recognition of two sources of variation: time series variation and sampling variation. They obtain some theoretical results regarding design consistency of the time series estimators, and uncorrelatedness of the signal and sampling error series. They also observe that the use of signal extraction results from time series analysis can improve survey estimates by reducing their mean square error.

For repeated surveys, better small area estimates can be obtained by combining the usual approach based on synthetic estimation with the use of time series models. Pfeffermann and Burck examine the statistical properties of such predictors. They illustrate the procedure with the use of data on home sale prices.

Time series described by ARIMA regression models with survey errors following an ARMA process is the subject of Binder and Dick's paper. Such models can be applied to data from surveys with a two-stage design where the first stage units are replaced randomly, while the second stage units have a rotating panel design. The authors give an example using Labour Force Survey data.

Brillinger studies the relationship of births to time and geography using data for women aged 25-29 in Saskatchewan. Smooth surfaces are obtained from data aggregated by census division. The Poisson-lognormal distribution is also fitted to the data.

In the last paper of the special section, Laniel and Fyfe describe the problem of benchmarking sub-annual series and briefly review some solutions proposed in the literature. They then present two new methods - one based on a model for trends and the other on a model for levels - and discuss their suitability.

In his paper, Bandyopadhyay proves that for a class of estimators and sampling schemes, one can ignore the sampling weights when estimating a ratio. He applies this to a well-known example to illustrate the result and makes a comparison with estimation using a ratio of Horvitz-Thompson estimators.

In repeated surveys with rotation panels, knowledge of panel corrélations is essential for certain statistical analyses, such as studies of composite estimators. Lee provides methodology for estimating correlations between panel estimates in the Canadian Labour Force Survey.

Misdating or "telescoping" is a recognized source of errors in retrospective surveys. Silberstein estimates telescoping effects to obtain estimates for the unbounded first wave in the U.S. Consumer Expenditure Interview Survey. She finds that estimates from the first wave are greater than estimates from subsequent waves even after accounting for telescoping effects and concludes that a shorter recall period for the first wave improves reporting in subsequent waves.

Stasny presents several models for gross flows in the presence of nonresponse. The models are divided into those with symmetric and asymmetric transition probabilities. Methods for obtaining parameter estimates for the various models are developed and applied to victimization data from the U.S. National Crime Survey.

Finally, readers will notice that, with this issue, Survey Methodology has a new cover. The previous cover was used since December 1984 (Vol. 10 No. 2). Statistics Canada is making similar changes to all its publications to incorporate a unique logo and to create a standardized corporate look.

The Editor

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 167–180
Statistics Canada

Analysis of Repeated Surveys

WAYNE A. FULLER[1]

ABSTRACT

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 in which 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 survey designs constructed to enable estimation of the parameters of the measurement error process can be very efficient.

KEY WORDS: Survey sampling; Least squares; Measurement error; Gross change.

1. INTRODUCTION

There is considerable interest in the analysis of surveys that are repeated in time. Evidence of this interest is the recently published proceedings of a conference on panel surveys edited by Kasprzyk, Duncan, Kalton and Singh (1989), sessions at the meetings of the International Statistical Institute held in 1987 and 1989, and the Statistics Canada Symposium on Analysis of Data in Time held in October 1989. 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 the study of 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.

[1] Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.

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 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 relative to a one-time survey. The quality repetition of a survey requires maintaining consistent field, processing, data management, and estimation procedures over time. It is difficult to maintain cooperation over time and it is difficult to trace people who move. Response error is present in all surveys, but repeated surveys encounter problems of "conditioning" associated with repeated interviews. Also, response errors introduce inconsistencies into data collected over time. 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 square 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), these 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 a 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 longitudinal 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 Y plus the least squares estimator of Z is the least squares estimator of $Y + Z$. 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 because all available information cannot be used in the estimation. First, it is not possible to incorporate all data from the surveys of preceding times into a least squares analysis for the current time because the number of variables often exceeds the number of observations. Second, the releasing organization may be restricted in the number of times they can 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. A 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 the categories in 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.

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

Table 1

Hypothetical proportions for two points in time

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

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)

$P_{E\cdot 1}$	$P_{E\cdot 2}$	P_{EE}	$P_{\cdot E2}$	$P_{\cdot E3}$
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

Table 3

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

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

were observed at the first time and one half are new to the sample. We take as our vector of observations the vector containing the proportion of elements in category 1 in the one half of the sample that is not observed the second time [denoted by $P_{E\cdot 1}$], the proportion of elements in category 1 at time 1 in the remaining half of the sample [denoted by $P_{E\cdot 2}$], the proportion of 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 $P_{\cdot E2}$], 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 $P_{\cdot E3}$].

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 estimators. 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 elements, and to estimate the number in the first category at time 2, 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 proportion of the elements in category 1 at both time 1 and time 2, and the 200 elements observed at time 2 are used to estimate the time 2 proportion.

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

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

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

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

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

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

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 an 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 than our example.

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 linear 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 Γ_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 employment 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 of P_{EE} 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 considered to be 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) conducted 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 1984-1985 with weights for these individuals, (b) the sample of individuals that responded 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 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, $0XX$, $X0X$, $XX0$, $X00$, $0X0$, $00X$, where X denotes response and 0 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 square estimators for the selected characteristics become control variables for a next stage of estimation. Using regression weighting methods, weights are constructed for the individuals that responded at all time periods. 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 the 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. The procedure is less than fully efficient for the control variables only because a limited amount of information is used in the generalized least squares procedure.

The strong advantage of the outlined 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 disadvantage is that estimates for particular points in time are less than fully efficient.

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 $\hat{\theta}$ is the vector of regression coefficients obtained in the regression of Y_i on X_i using the set of complete observations, and $(\bar{y},\bar{x})$ is the mean vector for the elements observed at every time period. 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 nonnegative. 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 characteristics, 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 nonfederal 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, are 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 1,500 new segments were selected in areas of rapid urban growth. Data were collected on about 280,000 points in 1987.

Table 4
Illustration of estimation procedure

1982	1987				TOTAL
	Cropland	Other	Urban	Roads	
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

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 3,100 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 5,530 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, $\hat{\mu}_X$ 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.

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 developed at Iowa State University. The theory for the small area estimation procedure is described 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. Simple 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. Proper variance estimation requires the use of double sampling formulas.

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), Poterba and Summers (1985), and Singh and Rao (1990). 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,

$$
\begin{aligned}
P\{\text{response} = E|\text{true} = E\} &= 1 - \alpha + \alpha P_E, \\
P\{\text{response} = U|\text{true} = E\} &= \alpha P_U, \\
P\{\text{response} = U|\text{true} = U\} &= 1 - \alpha + \alpha P_U, \\
P\{\text{response} = E|\text{true} = U\} &= \alpha P_E,
\end{aligned}
$$

where α is the parameter of the response mechanism. Under this model the expected value for the proportion employed at any point in time is the true proportion. A consistent estimator for 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

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)

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

$$\pi_{EE} = (0.90)^{-2}\{0.91 - 0.93(0.94)(0.19)\}$$
$$= 0.9184.$$

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 $P_{E\cdot}$ and $P_{\cdot E}$ 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}.

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 θ_{ij} denote the ij-th element of the 2×2 matrix of probabilities observed in the reinterview study. That is, θ_{ij} is the probability that an individual responds i on the first interview and j

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

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.

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.

6. SUMMARY AND CONCLUSIONS

We have reviewed some topics associated with the analysis of repeated data, without attempting a complete discussion of the topic. We have shown that procedures based upon least squares have the potential to provide large gains in efficiency. Because of size and timing considerations, it is not possible to include all available information in the construction of the least squares estimators. Thus, in practice, the statistician must choose a subset of variables to use in the construction of least squares weights. Estimation for a two-period survey conducted by the U.S. Soil Conservation Service was described.

We illustrated the large biases that measurement error can produce in longitudinal estimates such as gross changes estimates. We showed that measurement error methods exist that can be used to construct consistent estimators. The use of one fourth of the available resources to estimate the variance of the measurement error in order to use measurement error estimation methods can be justified.

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., and FULLER, W.A. (1973). An unbiased response model for analysis of categorical data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 202-207.

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.

COCHRAN, 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.

FULLER, W.A. (1986). Small area estimation as a measurement error problem. *Proceedings of the Conference on Survey Research Methods in Agriculture*, (Ed. D. Faulkenberry), American Statistical Association and NASS, U.S. Department of Agriculture, Washington, D.C.

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 Section on Social Statistics, American Statistical Association*, 242-257.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, 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*, Series 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 Section on Social Statistics, 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* (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 Society*, Series B, 12, 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, 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.

SINGH, A.C., and RAO, J.N.K. (1990). Adjustments for classification error in gross flows. Unpublished manuscript, Statistics Canada, Ottawa, Canada.

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.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 181–194
Statistics Canada

Sample Maintenance Based on Peano Keys

KIRK M. WOLTER and RACHEL M. HARTER[1]

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.

KEY WORDS: Recurring surveys; Sample maintenance; Changing population units; Peano key.

1. INTRODUCTION

We are concerned with recurring surveys conducted over time and the maintenance they require. Let $\mathfrak{U}_t$ 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 $\mathfrak{U}_0$ 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 (*e.g.*, see 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 units;
(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.

[1] Kirk M. Wolter and Rachel M. Harter, Statistical Research Department, A.C. Nielsen Company, Nielsen Plaza, Northbrook IL 60062, USA.

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
- 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). Also see the broad comments of Duncan and Kalton (1987) on household surveys and Colledge (1989) on business surveys.

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 establishments 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 selection 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 performance 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 $\Re^2$ to $\Re^1$ such that points in $\Re^2$ 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 $\Re^2$ 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 $X = X_k \ldots X_3 X_2 X_1$ and $Y = Y_k \ldots Y_3 Y_2 Y_1$ represent the longitude and latitude of an arbitrary point in k-digit binary form. Then, the corresponding Peano key is $P = X_k Y_k \ldots X_3 Y_3 X_2 Y_2 X_1 Y_1$. Also see figure 1 for an example for the case $k = 4$. Note how simple it is to calculate the value of P.

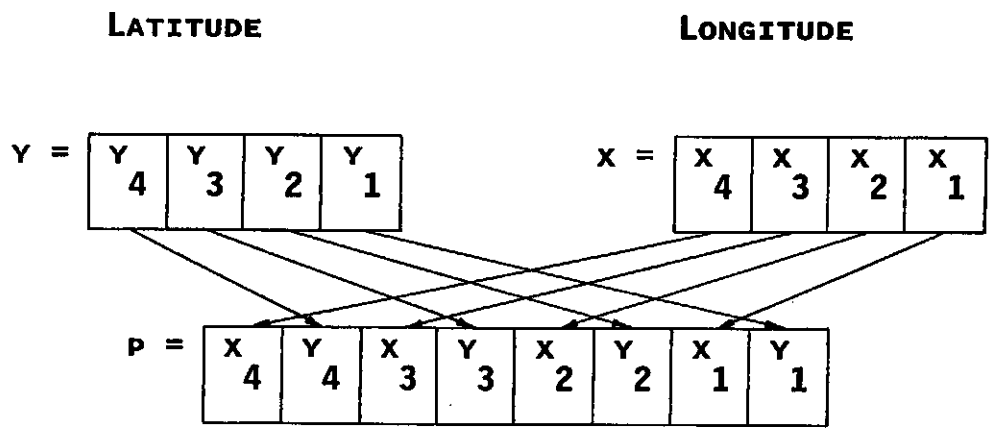


Figure 1. Creating the Peano Key by Bit Interleaving

Given k-digit (for any finite k) latitude longitude coordinates, the spacial "point" represented by the value of P is actually a square in $\mathfrak{R}^2$. 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 $\mathfrak{R}^2$.

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 $\mathfrak{R}^2$ to $\mathfrak{R}^1$.

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.

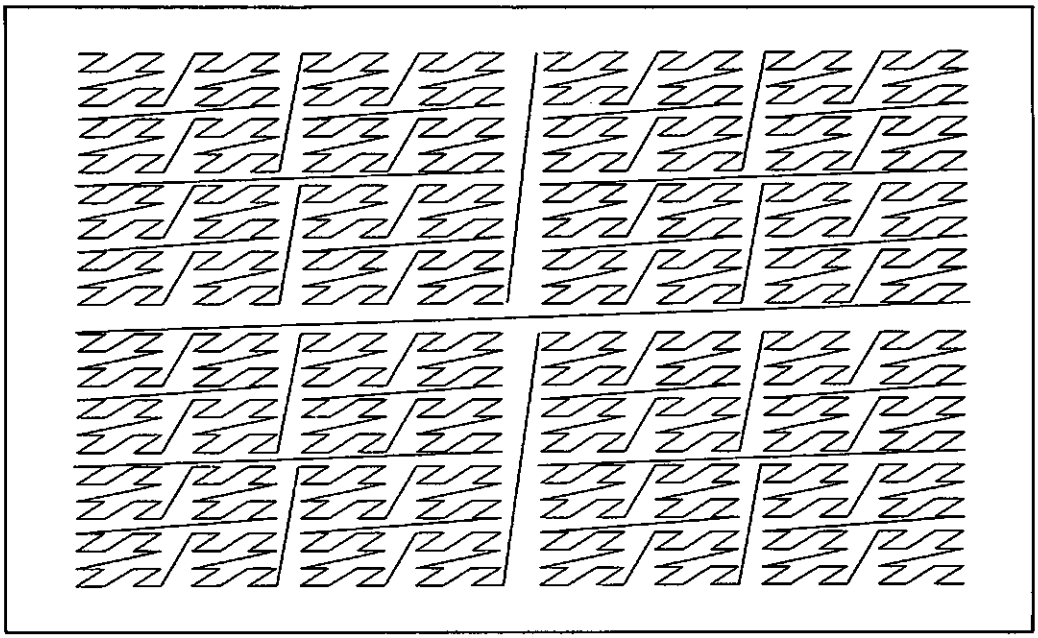


Figure 2. Peano Order Based on 1024 Points

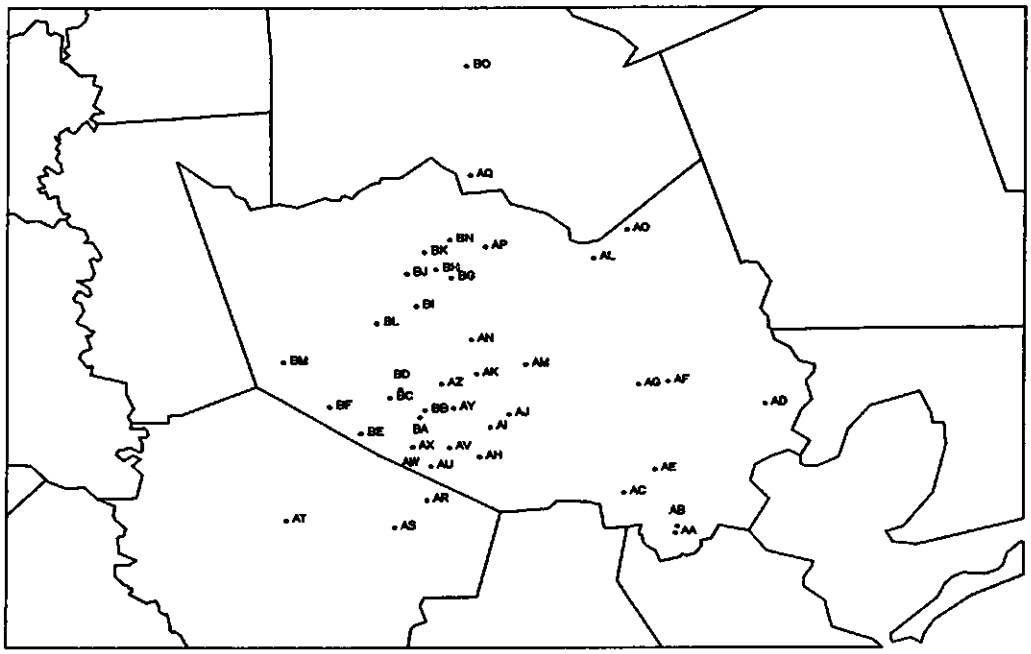


Figure 3. Chain of Retail Establishments in Peano Order

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 $\mathfrak{R}^2$ to $\mathfrak{R}^1$ 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 $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, 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 $n_i = n$. 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,

$$P_L \leq P_{11} < P_{21} < \ldots < P_{k1} < P_{12} < \ldots < P_{ij} < \ldots < P_{kn_k} \leq P_U.$$

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 $S_{kn_k} = [P_{kn_k}, P_U] \cup [P_L, P_{11})$ so that the entire Peano range $[P_L, P_U]$ 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 one 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 $U_{ij1}, U_{ij2}, \ldots, U_{ijB_{ij}}$ denote the stores in segment S_{ij}. The original store is now labeled U_{ij1}, whereas $U_{ij2}, U_{ij3}, \ldots, U_{ijB_{ij}}$ 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 $U_{ij1}, U_{ij2}, \ldots, U_{ijB_{ij}}$ the probabilities $p_{ij1}, p_{ij2}, \ldots, p_{ijB_{ij}}$, where $p_{ijb} > 0$ and $\sum p_{ijb} = 1$. Now choose one of the stores according to this probability measure. Subsampling is independant 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 measures 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

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

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

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

where

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

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. We 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 $\pi's$ 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. A replacement store will necessarily be selected within a given segment whenever the sample store from the segment has died -- except when there is a death but no birth and $B_{ij} = 0$ -- and a replacement store may be selected even when the sample store is alive and well.

In the exceptional case, where $B_{ij} = 0$, the sample size drops by 1. An interesting problem for future research is to investigate the mean square error of this rule versus that of an alternative rule which selects a replacement store from the same zone of k stores, instead of permitting the sample size to drop by 1. This alternative is conditionally unbiased but unconditionally biased.

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 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 Chronically Nonusable Stores or Scanning Conversions

In this final subsection, we present sample maintenance rules for handling stores that are chronically nonusable, such as stores that do not scan; do scan but with such poor discipline as to render their data faulty and nonusable; or refuse to participate in the survey. We shall explicitly discuss nonscanning stores and sample maintenance rules for handling conversions from nonscanning to scanning and vice versa, although the material that follows may be seen to apply more generally to all conditions of chronic nonusability. We shall let A denote the set of scanning stores and B the set of nonscanning stores, where $A \cup B$ spans the entire universe.

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 B to sampling. Sample only from the subuniverse of scanning stores A. As a given nonscanning store converts to scanning, then treat it as a birth, subjecting it to birth sampling. Prior to conversion, nonscanning stores B 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 to be committed to nonscanning stores.

To address case (ii), let 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

COLLEDGE, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.

DUNCAN, G.J., and KALTON, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97-117.

ERNST, L. (1989). Weighting Issues for Longitudinal Household and Family Estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.

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, 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*. New York: Springer Verlag.

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*.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 195–215
Statistics Canada

The Time Series Approach to Estimation for Repeated Surveys

WILLIAM R. BELL and STEVEN C. HILLMER[1]

ABSTRACT

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977) suggested the use of signal extraction results from time series analysis to improve estimates in repeated surveys, what we call the time series approach to estimation in repeated surveys. We review the underlying philosophy of this approach, pointing out that it stems from recognition of two sources of variation – time series variation and sampling variation – and that the approach can provide a unifying framework for other problems where the two sources of variation are present. We obtain some theoretical results for the time series approach regarding design consistency of the time series estimators, and uncorrelatedness of the signal and sampling error series. We observe that, from a design-based perspective, the time series approach trades some bias for a reduction in variance and a reduction in average mean squared error relative to classical survey estimators. We briefly discuss modeling to implement the time series approach, and then illustrate the approach by applying it to time series of retail sales of eating places and of drinking places from the U.S. Census Bureau's Retail Trade Survey.

KEY WORDS: Repeated surveys; Time series; Signal extraction; U.S. Retail Trade Survey.

1. INTRODUCTION

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977), hereafter SSJ, suggested the use of signal extraction results from time series analysis to improve estimates in repeated surveys. If the covariance structure of the usual survey estimates (Y_t) and their sampling errors (e_t) for a set of time points is known, these results produce the linear functions of the available Y_t's that have minimum mean squared error as estimators of the population values being estimated (say θ_t) for θ_t a stochastic time series. To apply these results in practice one estimates a time series model for the observed series Y_t and estimates the covariance structure of e_t over time using knowledge of the survey design.

Section 2 of this paper gives a brief overview of the basic results and framework for the time series approach. Section 3 considers some theoretical issues and section 4 some application considerations for the approach. In section 5 we illustrate the approach with an example using two time series from the Census Bureau's Retail Trade Survey.

2. BASIC IDEAS AND GENERAL CONSIDERATION OF THE TIME SERIES APPROACH

The basic idea in using time series techniques in survey estimation that distinguishes it from the classical approach is the recognition of *two sources of variability*. Classical survey estimation deals with the variability due to sampling – having not observed all the units in the population. Time series analysis deals with variability arising from the fact that a time series is not perfectly predictable (often linearly) from past data. Consider the decomposition:

[1] William R. Bell is Principal Researcher, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A., and Steven C. Hillmer is Professor, School of Business, University of Kansas, Lawrence, KS 66045, U.S.A.

$$Y_t = \theta_t + e_t, \tag{2.1}$$

where Y_t is a survey estimate at time t, θ_t is the population quantity of interest at time t, and e_t is the sampling error. The sampling variability of e_t is the focus of the classical survey sampling approach, which regards the θ_t's as fixed. From a time series perspective all three of Y_t, θ_t, and e_t can exhibit time series variation, as long as they are random and not perfectly predictable from past data. Standard time series analysis would treat Y_t directly and ignore the sampling error in the decomposition (2.1), not treating e_t explicity, but only indirectly in the aggregate Y_t. In fact, time series analysts typically behave as if the sampling variation is not present and the true values are actually observed. The most basic thing to keep in mind about the use of time series techniques in survey estimation is that there are two distinct sources of stochastic variation present that are conceptualized, modeled, and estimated differently.

2.1 Signal Extraction Results

Suppose that survey estimates Y_t are available at a set of time points labelled $t = 1, \ldots, T$. Let $\underset{\sim}{Y} = (Y_1, \ldots, Y_T)'$ and similarly define $\underset{\sim}{\theta}$ and $\underset{\sim}{e}$ so we have $\underset{\sim}{Y} = \underset{\sim}{\theta} + \underset{\sim}{e}$. Assuming the estimates Y_t are unbiased and θ_t and e_t are uncorrelated (see section 3.2)

$$E(\underset{\sim}{Y}) = E(\underset{\sim}{\theta}) \equiv \underset{\sim}{\mu} \equiv (\mu_1, \ldots, \mu_T)'$$

$$\Sigma_Y = \Sigma_\theta + \Sigma_e, \tag{2.2}$$

where E denotes expectation over both the sampling and time series model distributions, and Σ_Y is the covariance matrix of $\underset{\sim}{Y}$, *etc.* Here $\underset{\sim}{\mu}$ and Σ_θ refer to the time series structure of θ_t, which is not subject to sampling variation. If Y_t, θ_t, and e_t do not require differencing, it is well known that, since $\mathrm{Cov}(\underset{\sim}{\theta}, \underset{\sim}{Y}) = \Sigma_\theta$, using (2.2) the minimum mean squared error linear predictor of $\underset{\sim}{\theta}$ can be written

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{\mu} + \Sigma_\theta \Sigma_Y^{-1}(\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.3}$$

$$= \underset{\sim}{\mu} + \left(I - \Sigma_e \Sigma_Y^{-1}\right)(\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.4}$$

$$= \underset{\sim}{\mu} + \left(I + \Sigma_e \Sigma_\theta^{-1}\right)^{-1}(\underset{\sim}{Y} - \underset{\sim}{\mu}). \tag{2.5}$$

Another standard result is that the variance of the error of this estimate is

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_\theta - \Sigma_\theta \Sigma_Y^{-1} \Sigma_\theta = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e. \tag{2.6}$$

If normality of $(\underset{\sim}{\theta}, \underset{\sim}{Y})$ is assumed (2.3) - (2.5) give $E(\underset{\sim}{\theta} \mid \underset{\sim}{Y})$, the conditional expectation of $\underset{\sim}{\theta}$ given $\underset{\sim}{Y}$, and (2.6) gives $\mathrm{Var}(\underset{\sim}{\theta} \mid \underset{\sim}{Y})$, the conditional variance.

If Y_t requires differencing the preceeding results need to be modified. Assume e_t does not require differencing, but θ_t and Y_t need to be differenced once (*i.e.* by applying $1 - B$ where $BY_t = Y_{t-1}$). Let the differenced data be $W_t = (1 - B)Y_t = (1 - B)\theta_t + (1 - B)e_t$ for $t = 2, \ldots, T$. Let $\Delta = [\Delta_{ij}]$ be the $(T - 1) \times T$ differencing matrix with $\Delta_{ii} = -1$, $\Delta_{i,i+1} = 1$, and all other elements zero, and write $\Delta \underset{\sim}{Y} \equiv \underset{\sim}{W} = \Delta \underset{\sim}{S} + \Delta \underset{\sim}{e}$. Then we use

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{Y} - \hat{\underset{\sim}{e}} = \underset{\sim}{Y} - \Sigma_e \Delta' \Sigma_W^{-1} \Delta(\underset{\sim}{Y} - \underset{\sim}{\mu}), \tag{2.7}$$

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_e - \Sigma_e \Delta' \Sigma_W^{-1} \Delta \Sigma_e. \tag{2.8}$$

The expressions (2.7) and (2.8) also apply when θ_t and Y_t require a more general differencing operator (*e.g.* seasonal differencing), with appropriate definition of the differencing matrix Δ, as long as e_t does not require differencing. These results are analogous to (2.4) and (2.6), but with $\Delta' \Sigma_W^{-1} \Delta$ playing the role of Σ_Y^{-1}. The results are given in Bell and Hillmer (1990), where their optimality properties are discussed. They were essentially given by Jones (1980), but without real justification.

Scott and Smith (1974) and SSJ used classical signal extraction results equivalent to (2.3) - (2.6) based on covariance generating functions rather than covariance matrices. Bell (1984) considers such results for models involving differencing. Another approach (Binder and Dick 1989, Bell and Hillmer 1989) involves putting time series models for θ_t and e_t in state space form and using the Kalman filter and smoother, which can be viewed as an efficient way to compute the matrix results given above. Also, see Tam (1987) for use of the Kalman filter in an explicitly model-based approach to analysis in repeated surveys. In subsequent discussions we generally refer to the results (2.3) - (2.6), though our remarks easily extend to cover the use of (2.7) - (2.8).

In many cases, for time series Y_t and θ_t that are always positive, we will want to take logarithms of Y_t to help induce stationarity of θ_t and the sampling errors. In such cases we rewrite (2.1) as

$$Y_t = \theta_t(1 + \tilde{u}_t) = \theta_t u_t, \tag{2.9}$$

where $\tilde{u}_t = e_t/\theta_t$ and $u_t = 1 + \tilde{u}_t$. Taking logs we get

$$\log(Y_t) = \log(\theta_t) + \log(1 + \tilde{u}_t) = \log(\theta_t) + \log(u_t). \tag{2.10}$$

Letting $\underset{\sim}{\mu}$ and Σ_θ now refer to $\log(\underset{\sim}{\theta}) \equiv (\log(\theta_1), \ldots, \log(\theta_T))'$, and $\Sigma_Y = \Sigma_\theta + \Sigma_u$ refer to $\log(\underset{\sim}{Y})$, analogous to (2.4) our estimate is

$$\log(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{\mu} + \left[I - \Sigma_u \Sigma_Y^{-1}\right](\log(\underset{\sim}{Y}) - \underset{\sim}{\mu}). \tag{2.11}$$

The analogues to (2.6) - (2.8) are obvious. To estimate $\hat{\theta}_t$ we use $\exp[\log(\hat{\theta}_t)]$; alternatively, one could use $\exp[\log(\hat{\theta}_t) + \mathrm{Var}(\log(\hat{\theta}_t) - \log(\theta_t))/2]$ for a more "unbiased" estimate of θ_t with minimum mean squared error (see Granger and Newbold 1976).

Notice that (2.3) - (2.6) require knowledge of $\underset{\sim}{\mu}$ and any two of Σ_Y, Σ_θ, and Σ_e (the third can be obtained from (2.2)). In practice these will not be known exactly and will need to be estimated. Thus, the true minimum mean squared error linear predictor $\hat{\underset{\sim}{\theta}}$ cannot be obtained exactly and (2.6) or (2.8) understates the mean squared error (MSE) since it does not account for modeling errors. (See Binder and Dick (1989) and Eltinge and Fuller (1989).) The basic assumption underlying the application of the preceeding results, which we shall call the time series approach to survey estimation, is that $\underset{\sim}{\mu}$ and Σ_Y can be well-estimated from the time

series data on Y_t through a time series model, and Σ_e can be well-estimated using survey microdata and knowledge of the survey design (possibly also using a model). We discuss these issues further in section 4 and illustrate the approach with the example of section 5.

2.2 Some General Considerations of the Time Series Approach

Smith (1978), Jones (1980), and Binder and Dick (1986) review and discuss the approach known as Minimum Variance Linear Unbiased Estimation (MVLU). While both the MVLU and time series approaches can use data from time points other than t in estimating θ_t, they differ in that MVLU regards the θ_t's as fixed and still only treats one source of variation, that due to sampling. MVLU was developed for cases (such as many rotating panel surveys) where more than one direct estimate of θ_t is available for each t and the e_t's are correlated over time due to overlap in the survey design. The use of Y_j for $j \neq t$ in estimating θ_t then comes from generalized least squares results and the correlation of the e_t's. We can see the distinction in terms of our results for the simple case (2.1) where only one direct estimate, Y_t, of θ_t is available, by letting $\mathrm{Var}(\theta_t) \to \infty$ to get the MVLU. Then $\Sigma_\theta^{-1} \to 0$ and (2.5) becomes $\hat{\underset{\sim}{\theta}} = \underset{\sim}{Y}$, so without multiple estimates of θ_t the MVLU just uses Y_t to estimate θ_t. These remarks apply generally to composite estimation (Rao and Graham 1964, Wolter 1979), which is often used as an approximation to MVLU.

One question that may arise regarding the time series approach is why one should consider θ_t a stochastic time series? This issue has been discussed by SSJ and at length by Smith (1978). They observe that (1) users of data from repeated surveys treat the data Y_t as a stochastic time series in modeling and would do the same with θ_t if it were available (as it essentially is for surveys with very low levels of error), and (2) classical results (*e.g.* Patterson 1950) for estimation in repeated surveys (MVLU) assume a time series structure for the individual units in the population, while maintaining the anomalous position that θ_t, which is a function of these individual units (such as the total), is a sequence of fixed, unrelated quantities. In fact, if we assume θ_t is a sequence of fixed, unrelated quantities, then data through any time point are irrelevant to the future behavior of the true series θ_t. If this were the case, then there would be little point in doing the survey in the first place. The data would be out of date as soon as they were published. The real questions here are whether or not we can estimate the time series structure of θ_t and e_t well enough to make beneficial use of this in survey estimation, how worthwhile these benefits may be, and what risks are involved in doing so?

Along with opportunities for improving estimation in repeated surveys, the time series approach offers potential for improved results in other problems where typically only one of the two sources of variability is recognized. It also can potentially unify these as subproblems under one general approach. Such problems include preliminary estimation in repeated surveys (Rao, Srinath, and Quenneville 1989); seasonal adjustment (Wolter and Monsour 1981, Hausman and Watson 1985, Pfeffermann 1991); time series trend estimation and the related problem of detection of statistically significant change over time (Smith 1978); benchmarking, the reconciling of results from a repeated survey with the results from another survey or census estimating the same population characteristics (Hillmer and Trabelsi 1987, Trabelsi and Hillmer 1990); and inference about time series properties of the true series θ_t relevant to economic models (Bell and Wilcox 1990).

Finally, we note that the decomposition (2.1) or (2.10) does not allow for nonsampling errors, nor does the time series approach treat them explicitly. Whether nonsampling error is generally more or less of a problem for the time series approach than for the classical approach is unclear, but one may wish to consider the possible effects of known or suspected nonsampling errors on the time series estimators when applying them in particular situations.

3. THEORETICAL CONSIDERATIONS

We now obtain some theoretical results relevant to the time series approach, and some properties of the resulting estimators.

3.1 Consistency of Time Series Estimators

Following Fuller and Isaki (1981) we let Y_t^ℓ (from the ℓth sample at time t) be a sequence of estimators of the characteristic θ_t^ℓ of the ℓth population at time t where the populations and samples for $\ell = 1, 2, \ldots$ are nested. (See their paper for details.) Define $\underset{\sim}{Y}^\ell, \underset{\sim}{\theta}^\ell, \underset{\sim}{e}^\ell, \underset{\sim}{\mu}^\ell, \Sigma_y^\ell, \Sigma_\theta^\ell, \Sigma_e^\ell, \hat{\underset{\sim}{\theta}}^\ell$, and $\hat{\theta}_t^\ell$ in the obvious fashion. We consider what happens to the time series estimators $\hat{\underset{\sim}{\theta}}^\ell$ when the estimators $\underset{\sim}{Y}^\ell$ are consistent, *i.e.* $Y_t^\ell \to \theta_t^\ell$ in some fashion as $\ell \to \infty$ for $t = 1, \ldots, T$, with T, the length of the series, remaining fixed. For now we assume $\underset{\sim}{\mu}^\ell$, Σ_θ^ℓ, and Σ_e^ℓ are known for each ℓ, which generally means the time series models (including their parameter values) for the components are known. Since $\underset{\sim}{\mu}^\ell$ and Σ_θ^ℓ are really superpopulation parameters for the time series, θ_t^ℓ, we wish to estimate, we shall assume these are the same for each population ℓ, that is, $\underset{\sim}{\mu}^\ell \equiv \underset{\sim}{\mu}$ and $\Sigma_\theta^\ell \equiv \Sigma_\theta$ (a positive definite matrix) for all ℓ. This is also partly for convenience since we could get the same results assuming $\underset{\sim}{\mu}^\ell \to \underset{\sim}{\mu}$ and $\Sigma_\theta^\ell \to \Sigma_\theta$ as $\ell \to \infty$.

From (2.5) it would appear that $\underset{\sim}{Y}^\ell \to \underset{\sim}{\theta}^\ell$ would imply $\hat{\underset{\sim}{\theta}}^\ell \to \underset{\sim}{\theta}^\ell$ as long as $\Sigma_e^\ell \to 0$. This condition suggests we need mean square convergence of Y_t^ℓ to θ_t^ℓ. We thus consider estimators Y_t^ℓ of θ_t^ℓ such that $E[(Y_t^\ell - \theta_t^\ell)^2] = E[(e_t^\ell)^2] \to 0$ as $\ell \to \infty$. Since $E[(e_t^\ell)^2] = \text{Var}(e_t^\ell) + [E(e_t^\ell)]^2$ this implies both $\text{Var}(e_t^\ell) \to 0$ and $E(e_t^\ell) \to 0$. Assuming $Y_t^\ell \to \theta_t^\ell$ in mean square for $t = 1, \ldots, T$ thus implies $\Sigma_e^\ell \to 0$. We can now establish

Result 3.1: Consider $\hat{\underset{\sim}{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_T)'$ given by (2.4). If $Y_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for $t = 1, \ldots, T$, then $\hat{\theta}_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for $t = 1, \ldots, T$.

Proof: From $\underset{\sim}{Y}^\ell = \underset{\sim}{\theta}^\ell + \underset{\sim}{e}^\ell$ with $\Sigma_e^\ell \to 0$ we have $\Sigma_Y^\ell \to \Sigma_\theta$ (even if $\underset{\sim}{\theta}^\ell$ and $\underset{\sim}{e}^\ell$ are correlated.) From (2.4) we have

$$\hat{\underset{\sim}{\theta}}^\ell - \underset{\sim}{\theta}^\ell = (\underset{\sim}{Y}^\ell - \underset{\sim}{\theta}^\ell) - \Sigma_e^\ell(\Sigma_Y^\ell)^{-1}(\underset{\sim}{Y}^\ell - \underset{\sim}{\mu}). \tag{3.1}$$

The first term on the right converges to $\underset{\sim}{0}$ in mean square; the second has mean $\underset{\sim}{0}$ and variance $\Sigma_e^\ell(\Sigma_Y^\ell)^{-1}\Sigma_e^\ell \to 0$ as $\ell \to \infty$. Since both terms converge to $\underset{\sim}{0}$ in mean square so does $\hat{\underset{\sim}{\theta}}^\ell - \underset{\sim}{\theta}^\ell$.

Convergence in probability is a more familiar concept in survey sampling. If $Y_t^\ell \to \theta_t^\ell$ as $\ell \to \infty$ in probability for $t = 1, \ldots, T$ this does not guarantee $\Sigma_e^\ell \to 0$, which is mean square convergence, a stronger condition. If we assume there are random variables ζ_t with finite variance such that $|e_t^\ell| \le \zeta_t$ (almost surely) uniformly in ℓ, then $Y_t^\ell \to \theta_t^\ell$ in probability implies $Y_t^\ell \to \theta_t^\ell$ in mean square (Chung 1968, p. 64). Therefore, using Result 3.1, we have

Result 3.2: If $Y_t^\ell \to \theta_t^\ell$ in probability as $\ell \to \infty$ for $t = 1, \ldots, T$ and there exist random variables ζ_t with finite variance such that $|Y_t^\ell - \theta_t^\ell| \le \zeta_t$ (almost surely) uniformly in ℓ, then $\hat{\theta}_t^\ell \to \theta_t^\ell$ in probability as $\ell \to \infty$ for $t = 1, \ldots, T$.

These consistency results show that if the errors in the original estimates Y_t of θ_t are small (Σ_e is small) then the errors $\hat{\theta}_t - \theta_t$ will be small as well. From (3.1) we see this is because $\hat{\underset{\sim}{\theta}} - \underset{\sim}{Y}$ becomes small as Σ_e becomes small, thus when there is little error in the original estimates Y_t the time series approach will not change them much. Binder and Dick (1986) have noted this phenomenon, and also pointed out that in this case it does not matter what time series model is used. That is, the convergence to $\underset{\sim}{0}$ of (3.1) depends only on $\Sigma_e^\ell \to 0$ and not on $\underset{\sim}{\mu}$ or Σ_θ. Thus, the consistency results extend to allowing $\underset{\sim}{\mu}$, Σ_θ, and also Σ_e^ℓ to be replaced by estimates $\hat{\underset{\sim}{\mu}}^\ell$, $\hat{\Sigma}_\theta^\ell$, and $\hat{\Sigma}_e^\ell$ (which will generally come from estimated models – see sections 4 and 5), as long as $\hat{\underset{\sim}{\mu}}^\ell$ and $\hat{\Sigma}_\theta^\ell$ converge to something as $\ell \to \infty$ (it doesn't matter

what as long as the limit of $\hat{\Sigma}_\theta^\ell$ is positive definite) and $\hat{\Sigma}_e^\ell \to 0$, which should generally hold when $\Sigma_e^\ell \to 0$. It is also obvious that these results extend to the nonstationary case where $\hat{\underline{\theta}}$ is given by (2.7) instead of (2.4). While the results show that the time series estimates behave sensibly in the situation of small error in the original estimates Y_t, the gains from the time series approach will come in the opposite case - when $\mathrm{Var}(e_t)$ is large.

We can extend the consistency results to the case where we take logarithms and estimate $\log(\theta_t)$ using (2.11). In this case let $\Sigma_u^\ell = \mathrm{Var}(\log(\underline{u}^\ell))$ where $\log(\underline{u}^\ell) \equiv (\log(u_1^\ell), \ldots, \log(u_T^\ell))'$. If we are taking logarithms it is reasonable to assume Y_t^ℓ and θ_t^ℓ remain bounded away from 0, say $|Y_t^\ell| \geq \kappa$ and $|\theta_t^\ell| \geq \kappa$ (almost surely) for all t and ℓ for some constant $\kappa > 0$.

Result 3.3: If $Y_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for $t = 1, \ldots, T$, then $\log(Y_t^\ell) \to \log(\theta_t^\ell)$ and $\log(\hat{\theta}_t^\ell) \to \log(\theta_t^\ell)$ in mean square as $\ell \to \infty$ for $t = 1, \ldots, T$.

Proof: The analogue to (3.1) is

$$\log(\hat{\underline{\theta}}^\ell) - \log(\underline{\theta}^\ell) = (\log(\underline{Y}^\ell) - \log(\underline{\theta}^\ell)) - \Sigma_u^\ell(\Sigma_Y^\ell)^{-1}(\log(\underline{Y}^\ell) - \underline{\mu}).$$

If we can show $\Sigma_u^\ell \to 0$ we will have the result since this implies $\log(\underline{Y}^\ell) \to \log(\underline{\theta}^\ell)$ in mean square, and the second term on the right behaves exactly as that in (3.1). Notice

$$E[(\tilde{u}_t^\ell)^2] = E[(e_t^\ell)^2/(\theta_t^\ell)^2] \leq (E(e_t^\ell)^2)/\kappa^2 \to 0 \quad \text{as} \quad \ell \to \infty,$$

thus $E[(\tilde{u}_t^\ell)^2] = E[(u_t^\ell - 1)^2] \to 0$. This implies $\mathrm{Var}(u_t^\ell) \to 0$ and $E(u_t^\ell) \to 1$. By Jensen's inequality (Chung 1968, p. 45), since $\exp(\cdot)$ is a convex function,

$$1 \leq \exp(E[\log(u_t^\ell)^2]) \leq E(\exp[\log(u_t^\ell)^2]) = E[(u_t^\ell)^2].$$

But $E[(u_t^\ell)^2] = \mathrm{Var}(u_t^\ell) + [E(u_t^\ell)]^2 \to 1$ so $\exp(E[\log(u_t^\ell)^2]) \to 1$ implying $E[\log(u_t^\ell)^2] \to 0$. This yields $\mathrm{Var}(\log(u_t^\ell)) \to 0$ as desired.

As before we could get a convergence in probability result by imposing a boundedness condition on the $\log(u_t^\ell)$. Having $\log(\hat{\theta}_t)$ as an estimate of $\log(\theta_t)$, we have the following Corollary to Result 3.3 for using $\exp[\log(\hat{\theta}_t)]$ as an estimate of θ_t.

Corollary 3.4: If $Y_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for $t = 1, \ldots, T$, then (see (2.11)) $\exp[\log(\hat{\theta}_t^\ell)] \to \theta_t^\ell$ in probability as $\ell \to \infty$ for $t = 1, \ldots, T$.

Proof: Since $\log(\hat{\theta}_t^\ell) \to \log(\theta_t^\ell)$ in mean square implies convergence in probability, the result follows since $\exp(\cdot)$ is a continuous function (Chung 1968, p. 66).

An analogous result obviously holds for using $\exp[\log(\hat{\theta}_t^\ell) + \mathrm{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell))/2]$ to estimate θ_t, since then $\mathrm{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell)) \to 0$ as $\ell \to \infty$.

3.2 Uncorrelatedness of θ and e

Standard time series signal extraction results corresponding to (2.3) - (2.8) typically assume and θ_t and e_t are uncorrelated with each other at all leads and lags (equivalent to independence under normality). Previous papers on the time series approach to repeated survey estimation have merely assumed this, but since θ_t and e_t depend on the same population units it is not obvious that this assumption is valid. Fortunately, we can establish that it is valid under fairly general conditions. (Tam (1987) discusses how this fails under an explicitly model-based approach.)

We let y_{it} be the value of the characteristic of interest for the ith unit in the population at time t, and let $\Omega_t = \{y_{it}: i = 1, \ldots, N_t\}$ be the collection of all N_t of these units. We consider time points $t = 1, \ldots, T$ and let $\underset{\sim}{\Omega} = (\Omega_1, \ldots, \Omega_T)'$. The y_{it} are random variables, as is $\theta_t = \theta_t(\Omega_t)$, which is a function of the y_{it}. The sample at time t, s_t (denoting the indices, not the values, of the units selected), has probability of selection $p(s_t \mid \underset{\sim}{\Omega})$. The estimator Y_t of θ_t is a function of the values y_{it} for the units sampled, thus a function of both Ω_t and s_t, *i.e.* $Y_t = Y_t(\Omega_t, s_t)$. We could let Y_t depend on the sample at times other than t, but we ignore that here for simplicity.

We consider estimators Y_t of θ_t that are *design unbiased*, which we shall define as $E(Y_t \mid \underset{\sim}{\Omega}) \equiv \sum_{s_t} Y_t p(s_t \mid \underset{\sim}{\Omega}) = \theta_t$. We could alternatively define design unbiasedness as $E(Y_t \mid \Omega_t) \equiv \sum_{s_t} Y_t p(s_t \mid \Omega_t) = \theta_t$, and then would need to assume the sample selection process is such that $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t)$, so $E(Y_t \mid \underset{\sim}{\Omega}) = E(Y_t \mid \Omega_t)$. If the sample design is noninformative then s_t and $\underset{\sim}{\Omega}$ are independent, implying $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t) = p(s_t)$, and either definition of design unbiasedness reduces to $\sum_{s_t} Y_t p(s_t) = \theta_t$. This is the usual definition, which generally assumes the y_{it}, and so Ω_t and θ_t, are fixed. (The assumption $p(s_t \mid \underset{\sim}{\Omega}) = p(s_t \mid \Omega_t)$ allows the sample selection process at time t ($p(s_t \mid \underset{\sim}{\Omega})$) to depend on the population values at time t (Ω_t), but assumes the population values at time points other than t (Ω_j for $j \neq t$) offer no additional information on s_t beyond that in Ω_t. This might occur if sampling was with probability proportional to the size of an auxiliary variable at time t that was correlated with the y_{it} only at time t.) The assumptions we make here might even be generalized.

Result 3.5: If Y_t is design unbiased for all t then θ_t and e_t are uncorrelated time series.

Proof: Consider $\text{Cov}(\theta_t, e_j)$ for any two time points t and j. Since Y_j is design unbiased $E(e_j \mid \underset{\sim}{\Omega}) = E(Y_j - \theta_j \mid \underset{\sim}{\Omega}) = 0$, implying $E[E(e_j \mid \underset{\sim}{\Omega})] = E(e_j) = 0$. Also $E(\theta_t \cdot e_j \mid \underset{\sim}{\Omega}) = \theta_t \cdot E(e_j \mid \underset{\sim}{\Omega}) = 0$ implying $E(\theta_t \cdot e_j) = 0$. Thus $\text{Cov}(\theta_t, e_j) = E(\theta_t \cdot e_j) - E(\theta_t)E(e_j) = 0$.

Comment: If $E(e_j \mid \underset{\sim}{\Omega})$ does not depend on $\underset{\sim}{\Omega}$ then e_j is said to be "mean independent" of $\underset{\sim}{\Omega}$, which is known to be a stronger condition than e_j and $\underset{\sim}{\Omega}$ uncorrelated, though not as strong as stochastic independence (unless we have normality). This shows that actually we only need $E(e_t \mid \underset{\sim}{\Omega}) = E(Y_t \mid \underset{\sim}{\Omega}) - \theta_t$ to not depend on $\underset{\sim}{\Omega}$ for θ_t and e_t to be uncorrelated time series. This would cover cases where Y_t has a constant additive bias (not dependent on Ω_t) as an estimate of θ_t, or, using approximate Result 3.6 which follows, a constant percentage (multiplicative) bias.

We now consider the logarithmic decomposition (2.10) when the Y_t are design unbiased. We assume that $\tilde{u}_j$ is $O_p(r_\ell)$ where $r_\ell \to 0$ as $\ell \to \infty$ in the superpopulation framework of the previous section, omitting the superscript ℓ from random variables here for convenience. (See Wolter (1985, p. 222) for definition of the order in probability notation $O_p(r_\ell)$. For example, when estimating a population mean we would often have $\text{Var}(\tilde{u}_j) \leq K/n_{j\ell}$ where K is some constant and $n_{j\ell}$ is the sample size at time j in the ℓth population. Then $\tilde{u}_j = O_p(n_{j\ell}^{-.5})$ from Wolter (1985, theorem 6.2.1).) From a Taylor series linearization of $\log(u_j) = \log(1 + \tilde{u}_j)$ we have from Wolter (1985, theorem 6.2.2)

$$\log(u_j) = \tilde{u}_j + O_p(r_\ell^2). \tag{3.2}$$

Using this we obtain the following.

Result 3.6: If Y_t is design unbiased for all t and $\tilde{u}_j$ is $O_p(r_\ell)$, then to terms that are $O_p(r_\ell^3)$, $\log(\theta_t)$ and $\log(u_t)$ are uncorrelated time series.

Proof: From theorem 6.2.5 of Wolter (1985) $\mathrm{Cov}(\log(\theta_t), \log(u_j)) = \mathrm{Cov}(\log(\theta_t), \tilde{u}_j) + O_p(r_t^3)$. Notice $E(\tilde{u}_j \mid \underset{\sim}{\Omega}) = E(e_j \mid \underset{\sim}{\Omega})/\theta_j = 0$ implies $E(\tilde{u}_j) = 0$, and $E(\log(\theta_t)\tilde{u}_j \mid \underset{\sim}{\Omega}) = \log(\theta_t)E(\tilde{u}_j \mid \underset{\sim}{\Omega}) = 0$ implies $E(\log(\theta_t)\tilde{u}_j) = 0$, so $\mathrm{Cov}(\log(\theta_t), \tilde{u}_j) = 0$, establishing the result.

3.3 Design-Based Properties of Signal Extraction Estimates

Unconditionally, $\hat{\underset{\sim}{\theta}}$ in (2.3) is unbiased ($E(\hat{\underset{\sim}{\theta}}) = E(\underset{\sim}{\theta}) = \underset{\sim}{\mu}$) and has minimum MSE given by (2.6). It is easy to see that this is not the case when viewed from a design-based perspective. Suppose we begin with design-unbiased estimators $\underset{\sim}{Y}$, *i.e.* $E(\underset{\sim}{Y} \mid \underset{\sim}{\Omega}) = \underset{\sim}{\theta}$. From (2.2) and (2.4) we have $\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} = (I - \Sigma_e \Sigma_Y^{-1})\,\underset{\sim}{e} - \Sigma_e \Sigma_Y^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu})$. With some algebra, we can show the design bias, variance, and MSE of $\hat{\underset{\sim}{\theta}}$ are given by

$$
\begin{aligned}
E(\hat{\underset{\sim}{\theta}} \mid \underset{\sim}{\Omega}) - \underset{\sim}{\theta} &= -\Sigma_e \Sigma_Y^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu}),\\
\mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \mid \underset{\sim}{\Omega}) &= \Sigma_e - \Sigma_e \Sigma_Y^{-1}\Sigma_e - \Sigma_e \Sigma_Y^{-1}\Sigma_\theta \Sigma_Y^{-1}\Sigma_e,\\
E[(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta})(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta})' \mid \underset{\sim}{\Omega}] &= \Sigma_e - \Sigma_e \Sigma_Y^{-1}\Sigma_e\\
&\quad - \Sigma_e \Sigma_Y^{-1}\left[\Sigma_\theta - (\underset{\sim}{\theta} - \underset{\sim}{\mu})(\underset{\sim}{\theta} - \underset{\sim}{\mu})'\right]\Sigma_Y^{-1}\Sigma_e. \qquad (3.3)
\end{aligned}
$$

From a design-based perspective we see use of $\hat{\underset{\sim}{\theta}}$ trades bias for a reduction in variance, since $\Sigma_e - \mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \mid \underset{\sim}{\Omega})$ is a positive semidefinite matrix. Whether this reduces the conditional MSE (3.3) below Σ_e, the MSE of $\underset{\sim}{Y}$, depends on the last two terms in (3.3), and in turn on $\underset{\sim}{\theta}$. There can be particular realizations of $\underset{\sim}{\theta}$ for which the conditional MSE of $\hat{\underset{\sim}{\theta}}$ exceeds Σ_e, though on average signal extraction reduces the MSE by $\Sigma_e \Sigma_Y^{-1} \Sigma_e$, since the unconditional expectation of the bracketed term in (3.3) is zero. (Of course, (3.3) is unusable in practice since it depends on $\underset{\sim}{\theta}$.) Also, as noted earlier, modeling error will contribute additional MSE to $\hat{\underset{\sim}{\theta}}$, so another fundamental question, more difficult to answer (see Eltinge and Fuller 1989), is how the real unconditional MSE of $\hat{\underset{\sim}{\theta}}$ compares to Σ_e?

4. APPLICATION CONSIDERATIONS

Application of the time series approach to survey estimation requires estimation of the autocovariance structure of the sampling errors, estimation of the mean and autocovariance structure of the signal, and computation of the estimates $\hat{\theta}_t$ and $\mathrm{Var}(\hat{\theta}_t - \theta_t)$ as discussed in section 2. The first two generally involve use of time series models, and are discussed in some detail in Bell and Hillmer (1989). Here we make some general remarks. We assume the Y_t are design unbiased estimators of the θ_t. We illustrate application of the methods in the next section with two time series from the Census Bureau's Retail Trade Survey.

Sampling error autocovariances, $\mathrm{Cov}(e_t, e_{t+k})$, can be estimated in an analogous fashion to sampling variances, $\mathrm{Var}(e_t)$, which is done routinely and for which many methods are available. (See Wolter 1985.) In practice, there may be difficulties in linking survey microdata over time to directly estimate sampling error covariances. Nevertheless, in what follows we assume we have available such estimates $\widehat{\mathrm{Cov}}(e_t, e_{t+k})$ for some set of time points t and lags k. Unfortunately, if there is a substantial amount of sampling error present (the situation where time series methods can make a difference), such autocovariance estimates are likely to have high variances themselves. This suggests some sort of averaging to improve the autocovariance estimates.

First, if we assume e_t is covariance stationary, so $\mathrm{Cov}(e_t,e_{t+k}) \equiv \gamma_e(k)$ depends on k but not t, then each $\widehat{\mathrm{Cov}}(e_t,e_{t+k})$ is estimating $\gamma_e(k)$ and we can simply average them, *i.e.* take $\hat{\gamma}_e(k) = (T-k)^{-1}\sum_t \widehat{\mathrm{Cov}}(e_t,e_{t+k})$ if we have $\widehat{\mathrm{Cov}}(e_t,e_{t+k})$ for $t = 1, \ldots, T-k$. Alternatively, $\widehat{\mathrm{Corr}}(e_t,e_{t+k}) = \widehat{\mathrm{Cov}}(e_t,e_{t+k})/[\widehat{\mathrm{Var}}(e_t)\widehat{\mathrm{Var}}(e_{t+k})]^{.5}$ can be averaged over t to estimate $\mathrm{Corr}(e_t,e_{t+k})$, which also depends on k but not t, and the variance estimates can be averaged as before.

Now suppose we are assuming e_t is relative covariance stationary, so $\mathrm{Cov}(e_t/\theta_t, e_{t+k}/\theta_{t+k}) = \mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k}) \equiv \gamma_u(k)$ depends on k but not t. If $\tilde{u}_t$ is $O_p(r_t)$ for all t, as in section 3.2, then from (3.2) and theorem 6.2.5 of Wolter (1985), $\mathrm{Cov}(\log(u_t),\log(u_{t+k})) = \mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k}) + O_p(r_t^3) \approx \gamma_u(k)$. Taking $\widehat{\mathrm{Cov}}(e_t,e_{t+k})/(Y_tY_{t+k})$ as estimates of $\mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k})$, these can be averaged over t to estimate $\gamma_u(k)$. Alternatively, using corollary 5.1.5 of Fuller (1976) we can show that $\mathrm{Corr}(\log(u_t),\log(u_{t+k})) = \mathrm{Corr}(\tilde{u}_t,\tilde{u}_{t+k}) + O_p(r_t^3)$, and taking as estimates of $\rho_u(k) \equiv \mathrm{Corr}(\tilde{u}_t,\tilde{u}_{t+k}), \{\widehat{\mathrm{Cov}}(e_t,e_{t+k})/Y_tY_{t+k}\}/\{[\widehat{\mathrm{Var}}(e_t)\widehat{\mathrm{Var}}(e_{t+k})]^{.5}/Y_tY_{t+k}\} = \widehat{\mathrm{Corr}}(e_t,e_{t+k})$, we can average the estimated autocorrelations of e_t over t to estimate $\rho_u(k)$, which are approximately the autocorrelations of $\log(u_t)$. Relative variance estimates can be averaged as before.

Actually, the usual survey estimates of variances and autocovariances will be estimating $\mathrm{Var}(e_t \mid \underset{\sim}{\Omega})$ and $\mathrm{Cov}(e_t,e_{t+k} \mid \underset{\sim}{\Omega})$. These estimates may also be suitable as estimates of $\mathrm{Var}(e_t)$ and $\mathrm{Cov}(e_t,e_{t+k})$, *e.g.* if they make sense from a model-based perspective. If not, and if Y_t is design unbiased so $E(e_t \mid \underset{\sim}{\Omega}) = 0$, then averaging autocovariance estimates over time still makes sense. First, if e_t is assumed stationary, then $\gamma_e(k) \equiv \mathrm{Cov}(e_t,e_{t+k}) = E[\mathrm{Cov}(e_t,e_{t+k} \mid \underset{\sim}{\Omega})]$, so we can average estimates of $\mathrm{Cov}(e_t,e_{t+k} \mid \underset{\sim}{\Omega})$ to estimate $\gamma_e(k)$. Or if e_t is relative covariance stationary, then since $E(\tilde{u}_t \mid \underset{\sim}{\Omega}) = E(e_t \mid \underset{\sim}{\Omega})/\theta_t = 0, \gamma_u(k) \equiv \mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k}) = E[\mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k} \mid \underset{\sim}{\Omega})] = \mathrm{Cov}(\log(u_t),\log(u_{t+k})) + O_p(r_t^3)$, and estimates of $\mathrm{Cov}(\tilde{u}_t,\tilde{u}_{t+k} \mid \underset{\sim}{\Omega})$ can be averaged to estimate $\gamma_u(k)$. It is less clear how to justify averaging estimates of conditional (on $\underset{\sim}{\Omega}$) correlations, since $E[\mathrm{Corr}(e_t,e_{t+k} \mid \underset{\sim}{\Omega})] \neq \mathrm{Corr}(e_t,e_{t+k})$, though this may be true to a sufficient approximation. In general, approaches to estimation of sampling error autocovariance structures bear more investigation.

Given an estimate of the sampling error covariance structure, and using any relevant information about the design of the survey, we can attempt to determine a time series model and its parameters to closely reproduce this structure. This is illustrated in the example of section 5.

We now turn to developing a model for the signal, θ_t. Since the behavior of most published time series Y_t is dominated by their signals (otherwise, they would not be published), in developing models for signals θ_t we can draw on experience modeling time series Y_t without allowing for sampling error. Such experience suggests use of nonlinear transformations, differencing, and regression mean functions in the model for θ_t will be important. The logarithm is the most common nonlinear transformation used in time series, and taking $\log(Y_t)$ lets us model $\log(\theta_t)$ through (2.10), with consequences for the sampling error discussed above. The following remarks are given in terms of use of (2.1), but apply equally well to use of (2.10). While other transformations could be considered, they would not generally yield a convenient decomposition of transformed Y_t in terms of transformed θ_t and some sampling error. Choosing between taking logarithms or not transforming seems sufficient for modeling many series.

Assuming e_t has mean zero (implied by design unbiasedness) and does not require differencing, θ_t and Y_t will require the same differencing and have the same mean function. The mean function can often be modeled with a linear regression function, $\mu_t = \underset{\sim}{X}_t'\underset{\sim}{\beta}$, for some vector of regression variables $\underset{\sim}{X}_t$ and parameters $\underset{\sim}{\beta}$. We often use ARIMA

(autoregressive-integrated-moving average) models to account for the needed differencing and to explain the autocovariance structure of the differenced θ_t. A convenient approach to developing the θ_t model is to first model Y_t ignoring the sampling error, and then use a model with the same regression terms and ARIMA order for θ_t. The parameters of the θ_t model can then be estimated using the time series data for Y_t and the previously developed model for e_t, holding the parameters in the model for e_t fixed. Diagnostic checking may suggest modifications to the θ_t model. The final fitted model can then be used in the signal extraction estimation of θ_t. The model fitting and signal extraction computations are not trivial; Kalman filter/smoother algorithms are discussed in Bell and Hillmer (1989). These have been implemented in some software recently developed in cooperation with members of the time series staff of the Statistical Research Division of the Census Bureau. This software was used in the analysis of the next section.

5. EXAMPLE: U.S. RETAIL TRADE SURVEY - SALES OF EATING AND DRINKING PLACES

As an illustrative example we analyze time series of sales (in millions of dollars) of Eating Places and of Drinking Places, which are estimated in the monthly U.S. Retail Trade Survey. The Retail Trade Survey has a list panel of large businesses that are selected into the sample with certainty and report sales every month, and 3 rotating list panels of smaller businesses that are selected into the sample by stratified simple random sampling. There is also a rotating panel area sample covering companies not in the list universe. Quarterly, a sample of new firm births is introduced, and firm deaths as determined from activity checks are removed from the sample. The rotating panels report current month and previous month sales at intervals of 3 months for the list sample and 6 or 12 months for the area sample. Horvitz-Thompson (HT) estimates of current and previous months' sales are constructed; the resulting time series shall be denoted Y_t' and Y_{t-1}'. From these composite estimators are constructed as described in Wolter (1979). The final composite estimates will make up our time series Y_t. (While it might be interesting to instead analyze Y_t' and Y_{t-1}' directly, these estimates are not saved for a long enough period of time for seasonal time series modeling.) Sampling variances are estimated using the random group method (Wolter 1985, chapter 2) for the list sample with 16 random groups, and the collapsed stratum method for the area sample. Further information on the survey is given in Isaki *et al.* (1976), Wolter *et al.* (1976), Wolter (1979), Garrett, Detlefsen and Veum (1987), Bell and Wilcox (1990).

There are several complicating factors in the survey. The sample is redesigned and independently redrawn about every five years, with new samples having been introduced in September of 1977, and January of 1982 and 1987. This produces a break in the covariance structure of e_t every five years, which can be handled by the Kalman filter/smoother as discussed in Bell and Hillmer (1989). We shall use data from September, 1977 through December, 1986, so there is one redrawing of the sample near the middle of our series. When a new sample is introduced approximate MVLU estimates are used for the first three months before switching to the composite estimates (Wolter 1979). This introduces a transient effect into the sampling error autocorrelations that we shall ignore. Finally, the monthly estimates are benchmarked to annual totals estimated from an annual survey and from the economic census taken every five years. To avoid this complication we use data that are not benchmarked. The reader should be aware, however, that for this reason the data used here do not agree with published estimates.

Table 1

Sampling Error Correlations for Horvitz-Thompson Estimates

	Lag					
	4	8	12	16	20	24
Eating Places						
Averaged[1]	.72	.71	.79	.63	.65	.77
From (5.1)[2]	.75	.69	.81	.60	.53	.61
Drinking Places						
Averaged[1]	.70	.67	.78	.60	.60	.61
From (5.1)[2]	.72	.66	.80	.56	.50	.59
Number of Correlations Averaged	23	19	15	11	7	3
Weights Used in Determining $\hat{\phi}$'s	1	1	1	.5	0	0

[1] Raw estimates of Corr(e_t', e_j') and Corr(e_{t-1}', e_{j-1}') were available for all pairs of months from January, 1973 through March, 1975. Averages of the correlations for the lags shown were taken after applying Fisher's transformation, and the results then transformed back.

[2] Correlations are shown from model (5.1) for $m = 4$ with parameters $\hat{\phi}^4 = .604$, $\hat{\phi}_{12} = .723$ (Eating Places) and $\hat{\phi}^4 = .580$, $\hat{\phi}_{12} = .714$ (Drinking Places). These parameter values were determined to minimize the weighted sum of squared deviations of the correlations from model (5.1) and the averaged correlations using the weights shown. Lags 20 and 24 were not used (given zero weight) because of the small number of correlation estimates available at these lags.

5.1 Development of Sampling Error Models

Our first step will be to develop a model for the correlation structure of the sampling errors. Let us write $Y_t' = \theta_t + e_t'$ for the current month (t) HT estimate, and $Y_{t-1}' = \theta_{t-1} + e_{t-1}'$ for the previous month $(t - 1)$ HT estimate. We shall use the same models for e_t' and e_{t-1}'. Estimates of Corr(e_t', e_{t-1}') are extremely high - typically .98 or higher. While this is partly artificial (due to businesses reporting the same figure for current and previous month sales, and possibly due to the way missing values are imputed), in the absence of other information it is difficult to distinguish characteristics of e_t' from those of e_{t-1}'.

Since the three rotating panels in the survey are drawn (approximately) independently (Wolter 1979), auto- and cross-correlations for (e_t', e_{t-1}') should be nonzero only for lags that are multiples of 3. Estimates of such lag correlations can be averaged over time assuming correlation stationarity. While estimates of lag correlations are not regularly produced for the Retail Trade Survey, this was done as part of a special study using micro-data (random group totals) from the Retail Trade Survey sample for January, 1973 through March, 1975, albeit at a time when the survey had four rotating list panels. Lacking more recent data, we "averaged" the correlations at lags 4, 8, 12, 16, 20, and 24 for e_t' and e_{t-1}'. (This was actually done after applying Fisher's transformation $.5 \log((1 + r)/(1 - r))$, to make the distribution of the transformed correlations more symmetric, and then transforming the results back.) The results are shown in Table 1. They show fairly strong positive correlation in the sampling errors, and evidence of seasonality from the correlations at lag 12. A possible model given such data is

$$(1 - \phi^m B^m)(1 - \Phi B^{12}) e_t' = v_{1t}, \tag{5.1}$$

where $m = 4$ for the 4-panel survey, with the same model assumed for e_{t-1}' with $v_{2,t-1}$ replacing v_{1t}. (v_{1t} and $v_{2,t-1}$ are white noise with variance σ_v^2.)

A particularly convenient property of (5.1) is that if the sampling error in each panel would follow (5.1) with $m = 1$ if it were observed every month, then for any number m (that is a divisor of 12) of independent panels reporting successively, e_t' follows (5.1). This allows us to use the 4-panel survey results in Table 1 to estimate ϕ^4 and Φ, and (assuming $\phi > 0$) convert these to estimates of ϕ^3 and Φ, the parameters of the model for the current 3-panel survey. This was done by finding ϕ^4 and Φ to minimize the sum of squared deviations of the correlations from (5.1) with those of Table 1. (Lags 20 and 24 were dropped, and lag 16 given a weight of .5, due to the smaller number of correlation estimates that were averaged together at these higher lags.) This resulted in $\hat{\phi}^3 = .685$, $\hat{\Phi} = .723$ for Eating Places, and $\hat{\phi}^3 = .664$, $\hat{\Phi} = .714$ for Drinking Places. The resulting correlations for $m = 4$ from (5.1) are shown in Table 1, and may be compared to the averaged correlations. More formal statistical estimation procedures for ϕ^3 and Φ, as well as a possible test of model fit, could be considered. (We may pursue this later if sampling error autocorrelation estimates can be produced from more recent micro-data from the 3-panel survey.)

We make the further assumption that $\mathrm{Corr}(e_t', e_{t-1-k}'') = \rho\,\mathrm{Corr}(e_t', e_{t-k}')$ for all k. To justify this, note the population regression of e_{t-1-k}'' on e_{t-k}' is $\rho e_{t-k}' + \epsilon$, where if ϵ is not uncorrelated with e_t', at least it is certainly small since $\mathrm{Var}(\epsilon) = (1 - \rho^2)\mathrm{Var}(e_t')$ and ρ is very near 1. With this assumption (5.1) leads to the following bivariate model for (e_t', e_{t-1}''):

$$(1 - \phi^3 B^3)(1 - \Phi B^{12})\begin{bmatrix} e_t' \\ e_{t-1}'' \end{bmatrix} = \begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} \quad \mathrm{Var}\begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} = \sigma_v^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{5.2}$$

with $\rho = \mathrm{Corr}(v_{1t}, v_{2,t-1}) = \mathrm{Corr}(e_t', e_{t-1}'')$. Estimates of $\mathrm{Corr}(e_t', e_{t-1}'')$ are regularly produced and were available for 1982 through 1986. Averaging these (with Fisher's transformation) produced $\hat{\rho} = .985$ for Eating Places and $\hat{\rho} = .986$ for Drinking Places.

We can now use (5.2) to derive a model for the sampling error of the linear form of the composite estimator (Wolter 1979), which is given by

$$\begin{aligned} Y_t''' &= (1 - \beta)Y_t' + \beta(Y_{t-1}'' + Y_t' - Y_{t-1}'') && \text{(preliminary estimator)}, \\ Y_{t-1} &= (1 - \alpha)Y_{t-1}'' + \alpha Y_{t-1}''' && \text{(final estimator)}. \end{aligned} \tag{5.3}$$

In the (3-panel) retail trade survey, values of $\alpha = .8$, $\beta = .75$ are used. It is easily seen that (5.3) also holds for the sampling errors, *i.e.* with Y replaced by e. We can use the resulting relations to derive the following equation for e_t in terms of e_t' and e_{t-1}'':

$$(1 - .75B)e_t = .2e_t''' - .75e_{t-1}'' + .8e_t'. \tag{5.4}$$

Using (5.2) and (5.4) we then get

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})e_t = .2v_{2t} - .75v_{2,t-1} + .8v_{1t}. \tag{5.5}$$

The right hand side is a first order moving average process (Box and Jenkins 1976, p. 121) whose parameters can be determined given estimates of σ_v^2 and ρ. Thus, (5.5) would yield an ARMA model for e_t.

Rather than pursue this further, we shall instead make the rather strong assumption that a model of the same form holds for $\log(u_t)$ in $\log(Y_t) = \log(\theta_t) + \log(u_t)$, thus

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})\log(u_t) = (1 - \eta B)c_t. \tag{5.6}$$

Table 2
Coefficients of Variation (CV)[1] for Retail Sales Estimates

	Horvitz-Thompson	Final Composite[2]	Signal Extraction[3]	
	CV	CV	Low	High
Eating Places	.042	.025	.017	.023
Drinking Places	.088	.052	.032	.038

[1] CV = (Relative Variance)$^{.5}$.
[2] The values for the final composite estimator are obtained using models (5.7a,b).
[3] The values for signal extraction actually vary over time, being highest at the end of the series and lowest near the middle. We show the lowest and highest values, which are attained for both series in January 1982 (low) and December 1986 (high). The signal extraction variances are not symmetric in time because the sample redraw in January 1982 is not exactly at the center of the series.

We do this because estimates of sampling variance for these series are highly dependent on the level of the series; estimates of relative variance are much more stable over time. We also assume we can use estimates of relative variance and of ρ in determining η and σ_c^2. Estimates Y_t', Y_{t-1}', $\widehat{\mathrm{Var}}(e_t')$ and $\widehat{\mathrm{Var}}(e_{t-1}')$ were available for 1982 through 1986. The resulting relative variance estimates were used in the spirit of maximum likelihood estimation for the lognormal distribution – taking the average of the logs of the relative variance estimates, adding one half of the sample variance of the logged estimates to this, and exponentiating the results. (Merely averaging the relative variance estimates produced similar results.) This was done separately for Rel Var(Y_t') and Rel Var(Y_{t-1}'), and these two results were then averaged, producing a common relative variance estimate that is constant over time. The results are shown in Table 2 under the heading "Horvitz-Thompson". Using these and the $\hat{\rho}$'s given earlier, one can solve for η and σ_c^2 for the right side of (5.6). The resulting sampling error models are

$$(1 - .75B)(1 - .685B^3)(1 - .723B^{12})\log(u_t) = (1 + .130B)c_t \qquad (5.7a)$$

$$\text{(Eating Places)} \qquad \hat{\sigma}_c^2 = 1.948 \times 10^{-5}$$

$$(1 - .75B)(1 - .664B^3)(1 - .714B^{12})\log(u_t) = (1 + .134B)c_t \qquad (5.7b)$$

$$\text{(Drinking Places)} \qquad \hat{\sigma}_c^2 = 9.301 \times 10^{-5}.$$

One can use the method of McLeod (1975,1977) to solve for Var(log(u_t)) in these models, which is an estimate of the relative variance of the final composite estimator. The results are shown in Table 2. The corresponding coefficients of variation, .025 for Eating Places and .052 for Drinking Places, are quite close to estimates published in the Census Bureau's Monthly Retail Trade Reports that are obtained more directly.

5.2 Time Series Modeling and Signal Extraction

Figures 1a,b show plots of the time series of final composite estimates Y_t for Eating Places and for Drinking Places, respectively. To develop models for θ_t we shall begin by modeling the Y_t series directly. Both series show trends and strong seasonality, with the magnitude of the seasonal fluctuations larger the higher the level of the series. This suggests taking logarithms and the need for differencing; both are typical for economic time series. Examination

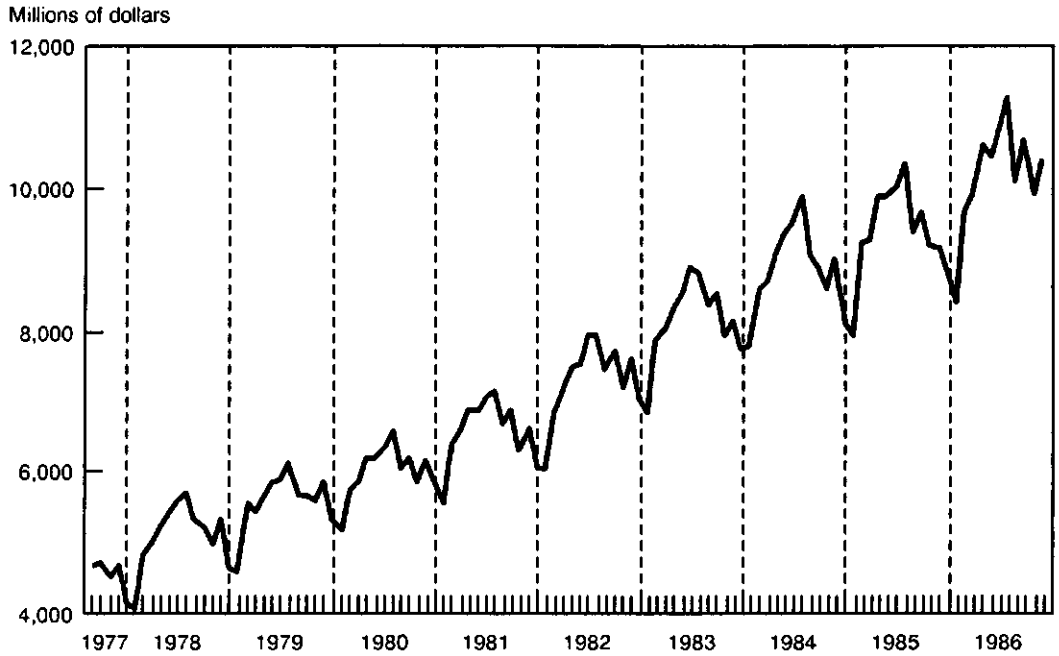


Figure 1.a Retail Sales of Eating Places – Composite Estimates (not benchmarked)

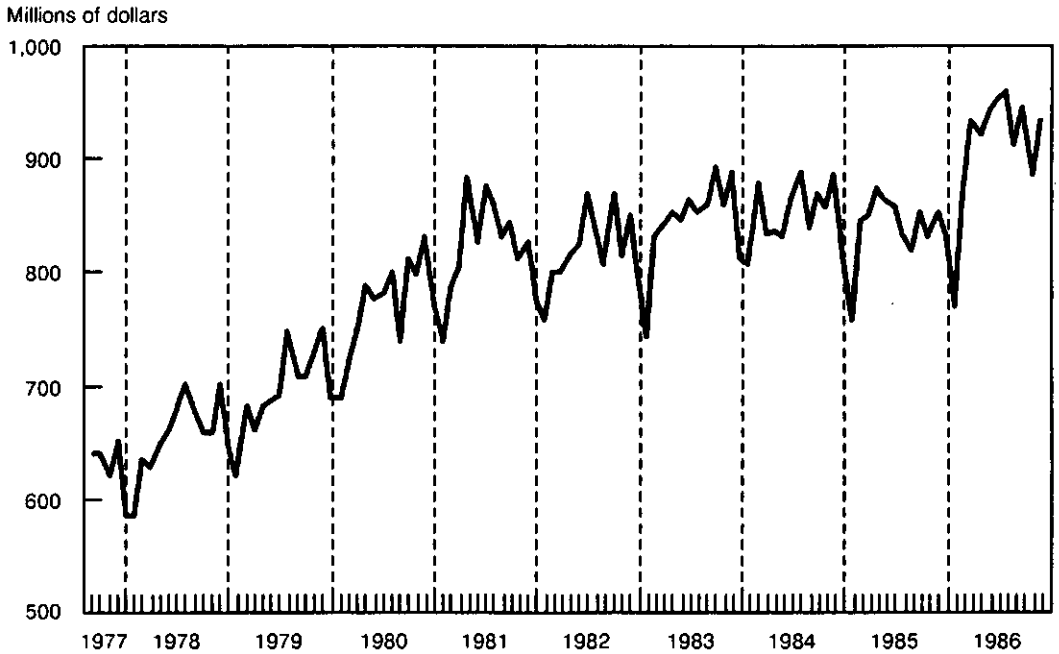


Figure 1.b Retail Sales of Drinking Places – Composite Estimates (not benchmarked)

of sample autocorrelations for $\log(Y_t)$ and its differences suggested the difference operator $(1 - B)(1 - B^{12})$ for both series. Retail trade series are known to contain trading-day variation, which can be modeled by including seven regression variables in the model: X_{1t} = number of Mondays in month t, ..., X_{7t} = number of Sundays in month t. Following Bell and Hillmer (1983), a more convenient parameterization is obtained by using instead the variables $T_{1t} = X_{1t} - X_{7t}$ (number of Mondays − number of Sundays), ..., $T_{6t} = X_{6t} - X_{7t}$ (number of Saturdays − number of Sundays), $T_{7t} = \sum_1^7 X_{it}$ (length of month t). To identify the ARMA structures, the autocorrelations and partial autocorrelations of the residuals from regressions of $(1 - B)(1 - B^{12}) \log(Y_t)$ on $(1 - B)(1 - B^{12})T_{it}$, $i = 1, \ldots, 7$, were examined. This suggested an ARIMA $(0,1,2)(0,1,1)_{12}$ model for Eating Places, and an ARIMA $(0,1,3)(0,1,1)_{12}$ model for Drinking Places. The resulting estimated models were

$$(1 - B)(1 - B^{12})\left[\log(Y_t) - \sum_i \beta_i T_{it}\right] = (1 - .25B - .22B^2)(1 - .79B^{12})\, a_t$$

$$\text{(Eating Places)} \qquad \hat{\sigma}_a^2 = .000230 \tag{5.8a}$$

$$(1 - B)(1 - B^{12})\left[\log(Y_t) - \sum_i \beta_i T_{it}\right] = (1 - .21B - .15B^2 + .03B^3)(1 - .56B^{12})\, a_t$$

$$\text{(Drinking Places)} \qquad \hat{\sigma}_a^2 = .000587. \tag{5.8b}$$

For brevity, we omit the estimates of the trading-day parameters. While the lag 2 and lag 3 moving average parameters in (5.8b) are small, we shall retain them since we shall only use (5.8a,b) as starting points for modeling $\log(\theta_t)$ for both series.

Taking models of the form of (5.8a,b) for $\log(\theta_t)$ with models (5.7a,b) for $\log(u_t)$, the parameters of the models for $\log(\theta_t)$ were estimated. For both series the seasonal moving average parameters were estimated to be very near 1(.985 for Eating Places and .992 for Drinking Places), implying nearly deterministic seasonality that can be modeled by cancelling a $(1 - B^{12})$ from both sides of the θ_t model and instead including a trend constant and a seasonal regression function of the form $\sum_1^{11} \gamma_i M_{it}$, where M_{1t} is 1 in January, −1 in December, and 0 otherwise, ..., M_{11t} is 1 in November, −1 in December, and 0 otherwise (Bell 1987). Estimation of the resulting models produced the following:

$$(1 - B)\left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it}\right] = .00762 + (1 - .20B - .29B^2)b_t$$

$$\text{(Eating Places)} \qquad \hat{\sigma}_b^2 = .000139 \tag{5.9a}$$

$$(1 - B)\left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it}\right] = .00352 + (1 - .18B - .09B^2 - .42B^3)b_t$$

$$\text{(Drinking Places)} \qquad \hat{\sigma}_b^2 = .000244. \tag{5.9b}$$

We again omit the estimates of the regression parameters. We do not provide standard errors for the ARMA parameters; doing so for models of the sort used here is a topic for further research, made particularly difficult here by the unrealistic assumption that the sampling error

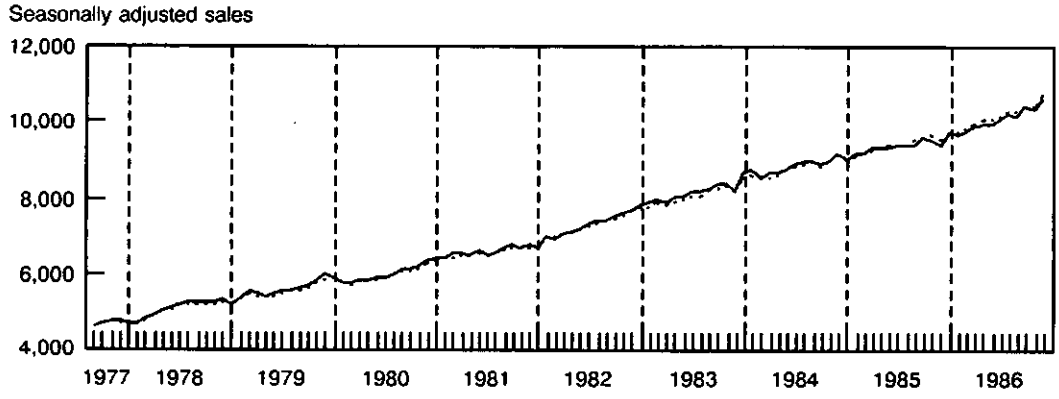


Figure 2.a Eating Places: Composite (solid) and Signal Extraction (dotted) Estimates

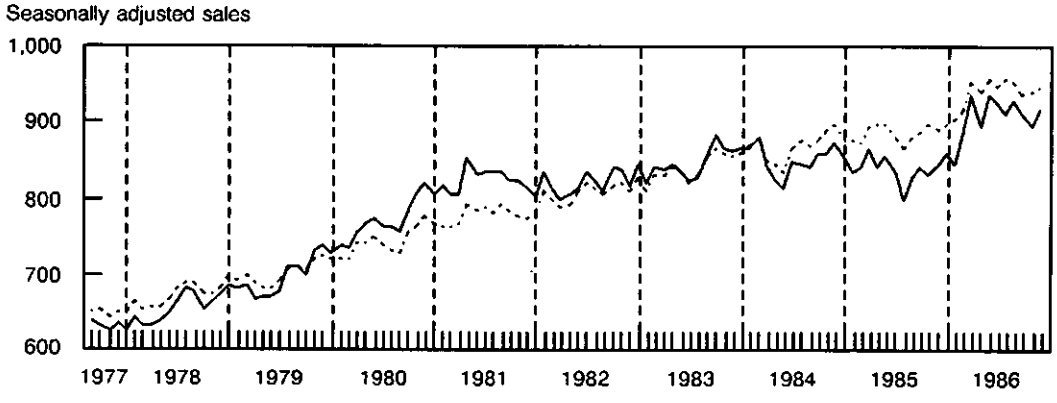


Figure 2.b Drinking Places: Composite (solid) and Signal Extraction (dotted) Estimates

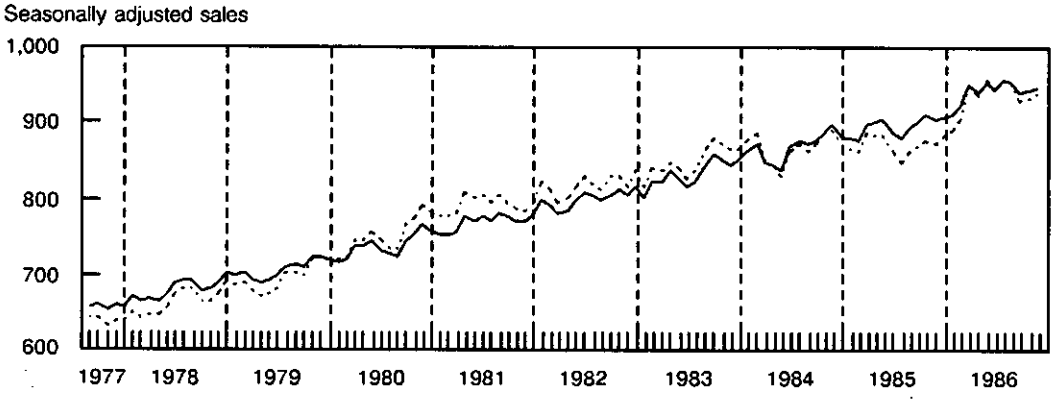


Figure 2.c Drinking Places: Alternative Signal Extraction Estimates

model is known. Examination of standardized residuals produced by the Kalman filter, and of their autocorrelations, suggested no major inadequacies with the fitted models for either series.

The estimated models, (5.7a,b) with (5.9a,b), were used to produce signal extraction estimates of $\log(\theta_t)$, which were then exponentiated to produce estimates of θ_t. The results are shown in Figures 2a,b for the series with the estimated seasonal and trading-day effects removed. Notice that signal extraction makes only slight differences in the estimates for Eating Places,which contained little sampling error (low relative variance), but it makes a considerable difference in the estimates for Drinking Places, which contained much more sampling error (higher relative variance). Signal extraction variances for $\log(\theta_t)$ were also produced; these are relative variances for the estimates of θ_t. Table 2 shows that, depending on the location in the series, signal extraction produces about an 8%-32% improvement in CV over the final composite estimates for Eating Places (though the composite estimate CV is small), and about a 27%-38% improvement in CV for Drinking Places. As noted previously, these results are optimistic, since they assume the true component models are those that were estimated. To partly address concerns about this, we next examine the sensitivity of the results for Drinking Places to variation in the model parameters.

5.3 Sensitivity Analysis for Drinking Places

Here we focus on sensitivity of results to variation in the sampling error model, since this was determined with less information than the signal model. Our approach is to vary parameters of the sampling error model, then reestimate the signal model and redo the signal extraction. While it would be preferrable to have more formal statistical measures of the signal extraction error due to model error (which the present state of theory and computer software does not allow), this approach should at least help indicate in what respects the signal extraction results are sensitive to parameter variation and in what respects they are not.

Comparing models (5.8b) and (5.9b) gives some indication of the sensitivity of the signal model to changes in σ_c^2, the innovation variance of the sampling error model, since (5.8b) corresponds to $\sigma_c^2 = 0$ and (5.9b) to $\sigma_c^2 = 9.3 \times 10^{-5}$. The most noticeable differences are in the estimate of σ_b^2, which is to be expected, and in the estimate of the seasonal moving average parameter, η_{12} say, which was found to be essentially 1 in obtaining (5.9b). Reestimation of the signal model for other values of σ_c^2 yielded $\hat{\eta}_{12} \geq .99$ as long as $\sigma_c^2 \geq 3.0 \times 10^{-5}$. In light of this, and to simplify presentation of results, we assume $\eta_{12} = 1$ and use a signal model with seasonal indicator variables as in (5.9b).

Figure 2.c. shows (seasonally and trading-day adjusted) signal extraction estimates $\hat{\theta}_t$ corresponding to sampling error models with $(\phi^3,\Phi) = (.564,.614)$ and $(.764,.814)$, and with $\rho = .986$ and $\mathrm{Var}(\log(u_t)) = .00776$ (the relative variance of the Horvitz-Thompson estimates) held fixed. These cover the extremes of $\hat{\theta}_t$ for the sensitivity analysis. The nature of the different estimates $\hat{\theta}_t$ we have generated seems to roughly correspond to the value of $\mathrm{CV}_{56} = [\mathrm{Var}(\log(\hat{\theta}_{56}) - \log(\theta_{56})]^{1/2}$, the signal extraction coefficient of variation achieved at the middle of the series. (CV_{56} is very close to the lowest value, which is achieved at $t = 53$ - see Table 2.) The lower CV_{56} is, the smoother $\hat{\theta}_t$ is. CV_{56} is 2.78%, 3.28%, and 3.70% for (ϕ^3,Φ) equal to $(.564,.614)$, $(.664,.714)$, and $(.764,.814)$ respectively. Other estimates $\hat{\theta}_t$ we generated lie closest to the signal extraction estimate in Figure 2.b. or 2.c. with the closest CV_{56}.

We now consider the sensitivity of CV_{56} to variations in the sampling error model parameters, beginning with ρ. The only parameter in (5.7b) affected by a change in ρ is η. Table 3 reports the values of η and corresponding values of ρ considered, and the resulting CV_{56}'s. We see CV_{56} is somewhat sensitive to changes in ρ, especially increases: CV_{56} for $\rho = 1$ (3.49) is 6% larger than for $\rho = .985$ (3.28), the value used for (5.7b).

Table 3

Sensitivity of CV_{56}[1] for Drinking Places to Changes in η (Changes in ρ)

η	.00	−.05	−.10	−.15	−.20	−.25
ρ	.9375	.9642	.9792	.9888	.9953	1.000
CV_{56}	3.03	3.12	3.21	3.31	3.40	3.49

[1] CV_{56} is the signal extraction coefficient of variation for $t = 56$ (the middle of the series), expressed as a percentage, *i.e.* the square root of $\mathrm{Var}(\log(\hat{\theta}_t) - \log(\theta_t))$ multiplied by 100.

Table 4

Sensitivity of CV_{56} for Drinking Places to Changes in $\mathrm{Var}(\log(u_t))$[1] (Changes in σ_c^2)

$\mathrm{Var}(\log(u_t))$	.00676	.00726	.00776	.00826	.00876
CV(HT)[2]	8.22	8.52	8.81	9.09	9.36
$\sigma_c^2 \times 10^5$	8.16	8.76	9.30	9.97	10.57
CV_{56}	3.16	3.23	3.28	3.35	3.40

[1] $\mathrm{Var}(\log(u_t))$ is the relative variance of the Horvitz-Thompson estimators.
[2] CV(HT) is the coefficient of variation of the Horvitz-Thompson estimators, expressed as a percentage, *i.e.* the square root of $\mathrm{Var}(\log(u_t))$ multiplied by 100.

Table 5

Sensitivity of Results for Drinking Places to Changes in (ϕ^3, Φ)

		(i) Values of $\sigma_c^2 \times 10^5$ for given (ϕ^3, Φ)				
		ϕ^3				
		.564	.614	.664	.714	.764
	.614	16.90	14.70	12.36	9.98	7.64
	.664	15.03	13.00	10.87	8.72	6.62
Φ	.714	13.04	11.23	9.30	7.44	5.60
	.764	10.96	9.40	7.78	6.15	4.58
	.814	8.79	7.51	6.17	4.85	3.58
		(ii) Values of CV_{56} for given (ϕ^3, Φ)				
		ϕ^3				
		.564	.614	.664	.714	.764
	.614	2.78	2.88	2.99	3.12	3.27
	.664	2.95	3.04	3.14	3.26	3.38
Φ	.714	3.10	3.19	3.28	3.39	3.50
	.764	3.24	3.33	3.42	3.51	3.60
	.814	3.36	3.45	3.54	3.62	3.70

We next consider the sensitivity of CV_{56} to changes in $\mathrm{Var}(\log(u_t))$. The only sampling error model parameter this affects is σ_c^2. Table 4 reports the values of $\mathrm{Var}(\log(u_t))$, its square root CV(HT), the corresponding σ_c^2, and the resulting CV_{56}. We see less sensitivity of CV_{56} here than in Table 3.

Finally, we examine the sensitivity of CV_{56} to ϕ^3 and Φ. Holding $\mathrm{Var}(\log(u_t))$ fixed at .00776 and changing (ϕ^3,Φ) also changes σ_c^2. Table 5 reports the grid of values used for (ϕ^3,Φ), and resulting values of σ_c^2 and CV_{56}. Notice σ_c^2 varies more here than in Table 4. We see CV_{56} increases substantially as ϕ^3 and Φ are increased.

We conclude from this analysis that moderate changes in the sampling error model parameters have relatively small impacts on $\hat{\theta}_t$. The largest changes we observed in $\hat{\theta}_t$ were around 2 percent. The same moderate changes in the sampling error model parameters have relatively larger impacts on the signal extraction variances, with CV_{56}'s changing by as much as 17 percent. This suggests that for this example the greatest concern in not knowing the sampling error model parameters may be in the effect on signal extraction variances, and the resulting measures of improvement over the composite estimates. However, in all the cases considered in the sensitivity analysis the signal extraction estimates showed a significant improvement in variance.

5.4 Conclusions

The Drinking Places example illustrates the potential gains that may be achieved with the time series approach to survey estimation. Both examples also illustrate the complex and delicate nature of the time series modeling that may be required. We view the results as preliminary for several reasons. First,the optimistic nature of the signal extraction variances that do not reflect parameter estimation error has been mentioned. Second, we have no clear explanation of why the signal extraction estimates lie above or below the composite estimates for long stretches of time. (This is obvious in Figure 2.b., and actually the case in Figure 2.a. as well.) For the Drinking Places example this behavior was evident throughout the sensitivity analysis, and so does not appear to be due to uncertainty in the parameters of the sampling error model. We are in the process of exploring whether this may be due to the forms of the sampling error model or signal model being incorrect. In fact, Bell and Wilcox (1990) report that the correlations of e_t' and e_{t-1}' at lags not multiples of three are not necessarily zero, as was assumed by the model.

ACKNOWLEDGEMENTS

This paper is based in part upon work supported by the National Science Foundation under grant SES 84-01460, "On-Site Research to Improve the Government-Generated Social Science Data Base," and by the U.S. Bureau of the Census and the University of Kansas under Joint Statistical Agreements 87-14 and 88-27. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Census Bureau, or the University of Kansas. We wish to thank Phillip Kott and David Findley for valuable suggestions regarding the consistency results, Abdelwahed Trabelsi for his support as an ASA/NSF/Census Research Associate, and an anonymous referee whose comments have markedly improved the paper. Thanks also to Ruth Detlefsen, Michael Shimberg, and Carol Veum of the Business Division of the Census Bureau for providing data from and information about the Retail Trade Survey. Finally, special thanks to James Bozik, Mark Otto, and Marian Pugh for their extensive work on developing the computer software used in this paper. Any errors in the analysis of the examples are the responsibility of the authors.

REFERENCES

BELL, W.R. (1984). Signal extraction for nonstationary time series. *Annals of Statistics*, 12, 646-664.

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.

BELL, W.R., and HILLMER, S.C. (1989). Modeling time series subject to sampling error. Research Report 89/01, Statistical Research Division, Bureau of the Census.

BELL, W.R., and HILLMER, S.C. (1990). A Matrix approach to signal extraction for nonstationary time series models. Submitted for publication.

BELL, W.R., and WILCOX, D.W. (1990). The effect of sampling error on the time series behavior of consumption data. Paper presented at the CRDE/Journal of Econometrics Conference on Seasonality in Econometric Models, Montreal, Canada, May 1990.

BINDER, D.A., and DICK, J.P. (1986). Modelling and estimation for repeated surveys. Statistics Canada Technical Report. Social Survey Methods Division.

BINDER, D.A., and DICK, J.P. (1989). Implications of survey designs for estimating seasonal ARIMA models. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.

BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

CHUNG, K.L. (1968). *A Course in Probability Theory*. New York: Harcourt, Brace and World, Inc.

ELTINGE, J.L., and FULLER, W.A. (1989). Time series random component models for sample surveys. Paper presented at the winter meeting of the American Statistical Association, San Diego, CA.

FULLER, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

FULLER, W.A., and ISAKI, C.T. (1981). Survey design under superpopulation models. *Current Topics in Survey Sampling*, (Eds. D. Krewski, R. Platek, and J.N.K. Rao). New York: Academic Press, 199-226.

GARRETT, J.K., DETLEFSEN, R.E., and VEUM, C.S. (1987). Recent sample revisions and related enhancements for business surveys of the U.S. Bureau of the Census. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 141-149.

GRANGER, C.W.J., and NEWBOLD, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society*, Series B, 38, 189-203.

HAUSMAN, J.A., and WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

HILLMER, S.C., and TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.

ISAKI, C.T., WOLTER, K.M., STURDEVANT, T.R., MONSOUR, N.J., and TRAGER, M.L. (1976). Sample redesign of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 90-98.

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

McLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

McLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 26,194.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, to appear.

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.

RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. *Panel Surveys*, (Eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, M.P. Singh). New York: Wiley, 457-479.

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.

SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri). New York: Academic Press, 201-216.

TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.

TRABELSI, A., and HILLMER, S.C. (1990). Benchmarking time series with reliable benchmarks. *Applied Statistics*, 39, to appear.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

WOLTER, K.M., ISAKI, C.T., STURDEVANT, T.R., MONSOUR, N.J., and MAYES, F.M. (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*, 99-109.

WOLTER, K.M., and MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (Eds. D. Krewski, R. Platek, and J.N.K. Rao). New York: Academic Press, 199-226.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 217–237
Statistics Canada

Robust Small Area Estimation Combining Time Series and Cross-Sectional Data

D. PFEFFERMANN and L. BURCK[1]

ABSTRACT

The common approach to small area estimation is to exploit the cross-sectional relationships of the data in an attempt to borrow information from one small area to assist in the estimation in others. However, in the case of repeated surveys, further gains in efficiency can be secured by modelling the time series properties of the data as well. We illustrate the idea by considering regression models with time varying, cross-sectionally correlated coefficients. The use of past relationships to estimate current means raises the question of how to protect against model breakdowns. We propose a modification which guarantees that the model dependent predictors of aggregates of the small area means coincide with the corresponding survey estimators and we explore the statistical properties of the modification. The proposed procedure is applied to data on home sale prices used for the computation of housing price indexes.

KEY WORDS: Kalman filter; Linear constraints; State-space models.

1. INTRODUCTION

Statistical Bureaus are often confronted with the demand to provide reliable estimators for small area means. The problem with the production of such estimators is that the sample sizes within those areas are usually too small to allow the use of direct survey estimators. As a result, new estimators have been proposed in recent years which combine auxiliary information (obtained from a census or administrative records) with the survey data obtained from all the small areas. The common feature of these estimators is that they can be structured in general as a linear combination of two components: a "synthetic estimator" of the form $\bar{X}_i'\hat{\beta}$ where $\bar{X}_i$ represents the average auxiliary information at the small area level and $\hat{\beta}$ is a vector of estimated regression coefficients; and a "correction factor" of the form $(\bar{y}_i - \bar{x}_i\hat{\beta})$ where $\bar{y}_i$ and $\bar{x}_i$ are the sample means of the target and the auxiliary variables. The correction factors are used to account for the variability of the small area means not explained by the auxiliary variables. The major difference between the various estimators is in the approach followed to determine the weights assigned to the two components in the linear combination, ranging from a "design based approach" (Särndal and Hidiroglou 1989) to "empirical Bayes" (Fay and Herriot 1979) and "mixed linear models" (Battese, Harter and Fuller 1989, Pfeffermann and Barnard 1991).

Very few studies are reported in the literature on the possible use of the time series relationships of the data to further increase the efficiency of the small area estimators. This is despite the fact that many of the small area estimators are derived from repeated surveys such as labour force surveys. The econometric literature contains a vast number of studies on the combined modelling of time series and cross-sectional data, see *e.g.* Rosenberg (1973b), Johnson (1977, 1980), Maddala (1977, Chapter 7), Dielman (1983) and Pfeffermann and Smith (1985) for reviews. However, none of these studies is directed to the problem of estimating (predicting) small area means from survey data. Fitting time series models to survey data has been considered

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905. L. Burck, Unit for Statistical Analysis, Central Bureau of Statistics, Jerusalem 91130.

in the context of estimating aggregate population means, see the review papers of Smith (1979) and Binder and Hidiroglou (1988) and the more recent articles by Binder and Dick (1989), Tiller (1989) and Pfeffermann (1991). But again, these methods are not in routine use mainly because the classical survey estimators of the aggregate means are often almost as efficient when the models hold and more robust when the models fail to hold.

The situation is clearly different when dealing with a small area estimation problem; it seems to us that for this kind of problem, the use of time series models can be of great advantage. Although the exact nature of the model to be used in a particular application is obviously 'data dependent', the class of models we consider in the next section is broad enough to apply to many, if not most of the small area estimation problems arising in practice. These models have the further advantage that their estimation is relatively simple. Estimation issues are discussed in Section 3.

The use of a model always raises the question of how to protect against possible model failures and this question becomes even more sensitive when considering the use of a model for the production of official statistics. In Section 4 we consider this issue and propose a modification to the model dependent predictors which guarantees that for aggregates of the small area means for which the direct survey estimators can be trusted, the modified model predictors coincide with the survey estimators. The statistical properties of the modified predictors are explored. We conclude the article in Section 5 with empirical results which illustrate the performance of the model with and without the proposed modification. The data used for the illustrations are the sale prices of homes in the city of Jerusalem during the months of September 1985 through November 1989. These data are used routinely by the Central Bureau of Statistics in Israel for the computation of housing price indexes.

2. REGRESSION WITH CROSS-SECTIONALLY AND TIME VARYING COEFFICIENTS

2.1 A General Class of Models

In what follows we denote by $\underset{\sim}{Y}_{tk}$ the $n_{tk} \times 1$ vector of observations on a target variable Y, pertaining to an area k at time t, $k = 1, \ldots, K$, $t = 1, 2, \ldots$. We assume for convenience that $n_{tk} \geq 1$ but as becomes evident later on, the model permits that some of the areas not be observed at certain times. Let X_{tk} define the corresponding $n_{tk} \times (p + 1)$ design matrix of the auxiliary variables with a vector of ones as its first column. In many applications, the same row vector $\underset{\sim}{x}'_{tk}$ of auxiliary values applies to all the Y values of a given time so that $X_{tk} = \underset{\sim}{1}_{n_{tk}} \underset{\sim}{x}'_{tk}$ where $\underset{\sim}{1}_{n_{tk}}$ is a column vector of ones of length n_{tk}. This is the case when the only available data are the small area survey estimators. Confidentiality as well as processing costs often preclude the use of micro data on individual survey respondents. The theory described in this article is not restricted to the availability of the micro data (see the example in Section 2.2) but data availability has an obvious effect on model specifications and precision of estimation.

The regression model holding in area k at time t is defined as

$$\underset{\sim}{Y}_{tk} = X_{tk}\underset{\sim}{\beta}_{tk} + \underset{\sim}{\epsilon}_{tk};\ E(\underset{\sim}{\epsilon}_{tk}) = 0,\ E(\underset{\sim}{\epsilon}_{tk}\underset{\sim}{\epsilon}'_{tk}) = \sigma_k^2 I_{n_{tk}} \tag{2.1}$$

where $\underset{\sim}{\beta}'_{tk} = (\beta_{tk0}, \beta_{tk1}, \ldots, \beta_{tkp})$.

We define the (superpopulation) mean of the target variable values in area k at time t to be

$$\Theta_{tk} = E(M_{tk} \mid \underset{\sim}{\beta}_{tk}) = \bar{\underset{\sim}{X}}_{tk}\underset{\sim}{\beta}_{tk} \tag{2.2}$$

where

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} \quad \text{and} \quad \bar{X}_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} x'_{tki}$$

with $i = 1, \ldots, N_{tk}$ indexing the population units. Obviously, when $x'_{tki} \equiv x'_{tk}$, then $\bar{X}_{tk} = x'_{tk}$.

Let $\hat{\beta}_{tk}$ define an estimator for β_{tk}. Then $\hat{\Theta}_{tk} = \bar{X}_{tk}\hat{\beta}_{tk}$ and

$$\hat{M}_{tk} = \frac{1}{N_{tk}}\left[\sum_{i=1}^{n_{tk}} Y_{tki} + \sum_{i=n_{tk}+1}^{N_{tk}} x'_{tki}\hat{\beta}_{tk}\right] = \hat{\Theta}_{tk} + \frac{1}{N_{tk}}\left(\sum_{i=1}^{n_{tk}} (Y_{tki} - x'_{tki}\hat{\beta}_{tk})\right)$$

implying that in the usual case of small sampling rates within the areas, $\hat{\Theta}_{tk}$ can also be considered as an estimator of the finite population mean M_{tk}. For this reason we no longer distinguish between the finite and superpopulation means.

The notable feature of (2.1) is that the coefficients β_{tk} are allowed to vary both cross-sectionally and over time. The following equations specify the variation of the coefficients over time:

$$\begin{bmatrix}\beta_{tkj}\\ \beta_{kj}\end{bmatrix} = T_j \begin{bmatrix}\beta_{t-1,kj}\\ \beta_{kj}\end{bmatrix} + \begin{bmatrix}1\\ 0\end{bmatrix}\eta_{tkj}, j = 0, \ldots, p \tag{2.3}$$

where we use the notation β_{kj}, $j = 0, 1, \ldots, p$, to define fixed coefficients which we interpret below, and T_j to define fixed (2×2) matrices and where the residuals $\{\eta_{tkj}\}$ satisfy

$$E(\eta_{tkj}) = 0, E(\eta_{tkj}\eta_{tk\ell}) = \delta_{j\ell}, E(\eta_{tkj}\eta_{t-d,k\ell}) = 0 \quad \text{for} \quad d > 0. \tag{2.4}$$

The implication of (2.4) is that residuals of different coefficients pertaining to the same time t are allowed to be correlated but the serial and cross serial correlations are assumed to be zero.

Next, we illustrate the use of (2.3) by considering some simple cases:

(a) $T_j = \left[\begin{smallmatrix}0 & 1\\ 0 & 1\end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{kj} + \eta_{tkj}$ so that β_{kj} represents, in this case, a common mean. This is the well known Random Coefficient Regression Model (Swamy 1971) which is often used in econometric applications. Obviously, by postulating, $\text{var}(\eta_{tkj}) = 0$, the model reduces to the case of a fixed regression coefficient over time.

(b) $T_j = \left[\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$ which is the familiar random walk model, see *e.g.* Cooley and Prescott (1976) and LaMotte and McWhorter (1977) for application of this model in econometric studies. In this case the coefficient β_{kj} is redundant and should be omitted so that $T_j \equiv 1$.

(c) $T_j = \left[\begin{smallmatrix}\rho, & 1-\rho\\ 0, & 1\end{smallmatrix}\right]$ implies the first order autoregressive relationship $(\beta_{tkj} - \beta_{kj}) = \rho(\beta_{t-1,kj} - \beta_{kj}) + \eta_{tkj}$ considered by Rosenberg (1973a).

(d) $T_j = \left[\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \beta_{kj} + \eta_{tkj}$ which defines a local approximation to a linear trend (Kitagawa and Gersch 1984). The coefficient β_{kj} represents, in this case, a fixed slope.

It should be emphasized that different matrices T_j can be used for different coefficients β_{tkj}. In fact, by defining $\underset{\sim}{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \beta_{k1}, \ldots, \beta_{tkp}, \beta_{kp})$; $\tilde{T} = \text{diag}[T_0, T_1, \ldots, T_p]$, a block diagonal matrix with T_j as the j-th block; $\tilde{G} = I_{p+1} \otimes \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ where I_{p+1} is the identity matrix of order $p + 1$ and $\otimes$ defines the Kronecker product and $\underset{\sim}{\eta}'_{tk} = (\eta_{tk0}, \eta_{tk1}, \ldots, \eta_{tkp})$, the combined model holding for the coefficients $\underset{\sim}{\beta}_{tk}$ can be written as

$$\underset{\sim}{\alpha}_{tk} = \tilde{T}\underset{\sim}{\alpha}_{t-1,k} + \tilde{G}\underset{\sim}{\eta}_{tk}; \quad E(\underset{\sim}{\eta}_{tk}) = \underset{\sim}{0}, E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{t-d,k}) = A_d\Delta \tag{2.5}$$

where $A_d = 1$ for $d = 0$ and $A_d = 0$ otherwise, and $\Delta = [\delta_{ij}]$ is defined by the variances and covariances δ_{ij} (equation 2.4).

The model defined by (2.5) specifies the variation of the regression coefficients of a specific area over time. The common approach to account for cross-sectional relationships between small area means is to allow for random small area effects which are time invariant $\{u_k\}$. The general model defined by (2.1) and (2.3) includes this case by writing $\underset{\sim}{Y}_{tk} = \underset{\sim}{1}_{n_{tk}}u_{tk} + X_{tk}\underset{\sim}{\beta}_{tk} + \underset{\sim}{\epsilon}_{tk} = X^*_{tk}\underset{\sim}{\beta}^*_{tk} + \underset{\sim}{\epsilon}_{tk}$, say, and specifying $u_{tk} = u_{t-1,k} + \eta_{tk}$ with $u_{ok} = 0$, $\text{var}(\eta_{1k}) = \sigma^2_\eta$ and $\text{var}(\eta_{tk}) = 0$ for $t > 1$ (compare with case (b) above). By assuming in addition the autoregressive relationship defined by case (c) for the intercept variable and fixing the other regression coefficients (case (a) with zero residual variances), the resulting model is similar to the model considered by Choudhry and Rao (1989) except that in their general formulation of the model the observation residuals of equation (2.1) are allowed to be serially correlated. Notice that equation (2.1) now contains two random "intercept terms" but the model is nonetheless identifiable. Choudhry and Rao assume that the only available data are the survey estimators so that the estimation of the serial correlations needs to be carried out externally, using the micro observations. Alternatively, a model accounting for the serial correlations can be postulated. Choudhry and Rao assume an AR(1) model in their study.

A more general way to account for the cross-sectional relationships between the small area means is to allow for non zero correlations between the residual terms η_{tkj} and η_{tmj} of the models specifying the time series variation of the regression coefficients β_{tkj} and β_{tmj} operating in areas k and m (equation 2.4). Often it is reasonable to assume that the correlations decay as the distance between the areas increases. This can be formulated as, $E(\eta_{tkj}, \eta_{tmj}) = \delta_{jj}\rho_j f_j(k,m)$, $k \neq m$, where $f_j(k,m)$ is a monotonic decreasing function of the distances $D(k,m)$. The case of geometrically decaying correlations is obtained by defining $f_j(k,m) = \rho_j^{|k-m|-1}$. The case of fixed correlations is obtained by specifying $f_j(k,m) \equiv 1$ and in what follows we consider this case only. Allowing for fixed cross-sectional correlations for all the regression coefficients can be formulated as

$$E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{tm}) = D(\Delta)\emptyset, \quad k \neq m \tag{2.6}$$

where $D(\Delta)$ is the diagonal matrix with the variances δ_{jj} on the main diagonal and $\emptyset$ is another diagonal matrix composed of the correlations ρ_j.

Before concluding this section we present the model defined by (2.1), (2.5) and (2.6) in a state-space form. Presenting the model in this form has important computational advantages.

Let $\underset{\sim}{Y}'_t = (\underset{\sim}{Y}'_{t1}, \ldots, \underset{\sim}{Y}'_{tK})$ represent the vector of observations of length $n_t = \sum_k n_{tk}$ for all the areas at time t and let $\underset{\sim}{\epsilon}'_t = (\underset{\sim}{\epsilon}'_{t1}, \ldots, \underset{\sim}{\epsilon}'_{tK})$ represent the corresponding regression residuals. Define $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, \underset{\sim}{x}_{tk1}, \underset{\sim}{0}_{ntk}, \ldots, \underset{\sim}{x}_{tkp}, \underset{\sim}{0}_{ntk}]$ where $\underset{\sim}{0}_{ntk}$ is a vector of zeroes of length n_{tk} and $\underset{\sim}{x}_{tkj}$ is the vector of values for the j-th auxiliary variable, $j = 1, \ldots, p$. Let Z_t be the block diagonal matrix composed of the matrices Z_{tk}. The matrix Z_t is of order $n_t \times [K \times 2 \times (p + 1)]$. Define also $\underset{\sim}{\alpha}'_t = (\underset{\sim}{\alpha}'_{t1}, \ldots, \underset{\sim}{\alpha}'_{tK})$, $\underset{\sim}{\eta}'_t = (\underset{\sim}{\eta}'_{t1}, \ldots, \underset{\sim}{\eta}'_{tK})$, $\Sigma_t = \text{Diag}[\sigma^2_1 \underset{\sim}{1}'_{nt1}, \ldots, \sigma^2_K 1'_{ntK}]$, $T = I_K \otimes \tilde{T}$, and $G = I_K \otimes \tilde{G}$.

Using this notation, the model defined by (2.1), (2.5) and (2.6) can be written compactly as

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\epsilon}_t;\ E(\underset{\sim}{\epsilon}_t) = \underset{\sim}{0}, E(\underset{\sim}{\epsilon}_t \underset{\sim}{\epsilon}_t') = \Sigma_t \tag{2.7}$$

$$\underset{\sim}{\alpha}_t = T\underset{\sim}{\alpha}_{t-1} + G\underset{\sim}{\eta}_t;\ E(\underset{\sim}{\eta}_t) = \underset{\sim}{0}, E(\underset{\sim}{\eta}_t \underset{\sim}{\eta}_t') = \Lambda, \tag{2.8}$$

where $\Lambda = [\Lambda_{k\ell}]$, $k, \ell = 1, \ldots, K$ with $\Lambda_{k\ell} = \Delta$ when $k = \ell$ and $\Lambda_{k\ell} = D(\Delta)\emptyset$ when $k \neq \ell$. The matrices $\Lambda_{k\ell}$ are $(p + 1) \times (p + 1)$.

The model defined by (2.7) and (2.8) conforms to the classical state-space formulation, see, *e.g.* Anderson and Moore (1979) and Harvey (1984). By this formulation, (2.7) is the observation equation and (2.8) is the state equation with $\underset{\sim}{\alpha}_t$ defining the state vector. The apparent advantage of restructuring the model in a state space form is that the vectors $\underset{\sim}{\alpha}_t$, and hence the population means Θ_{tk}, as well as the estimation error variances can be estimated conveniently by means of the Kalman filter. We discuss the use of the filter in sections 3 and 4.

2.2 Explicit Estimators of the Small Area Means

In order to illustrate how past and neighbouring data are used under the model to "strengthen" the small area estimators we consider the case where the same vector $\underset{\sim}{x}_{tk}$ of auxiliary values applies to all the units of a given area at a given time. In this case the observation equation can be formulated in terms of the sample means, *i.e.*

$$\bar{Y}_{tk} = \underset{\sim}{x}_{tk}' \underset{\sim}{\beta}_{tk} + \bar{\epsilon}_{tk};\ E(\bar{\epsilon}_{tk}) = 0,\ E(\bar{\epsilon}_{tk}^2) = \sigma_k^2/n_{tk},\ k = 1, \ldots, K. \tag{2.9}$$

Suppose that the regression coefficients follow a random walk (case (b) of equation 2.3) so that for area k

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj};\ E(\eta_{tkj}) = 0,\ E(\eta_{tkj}\eta_{tk\ell}) = \delta_{j\ell}, j, \ell = 1, \ldots, p \tag{2.10}$$

and for areas $k \neq m$,

$$E(\eta_{tkj}\eta_{tmj}) = \delta_{jj}\rho_j;\quad E(\eta_{tkj}\eta_{tm\ell}) = 0, j \neq \ell. \tag{2.11}$$

The random walk model implies that the coefficients drift slowly away from their initial value with no inherent tendency to return to a mean value. Obviously, for residuals η_{tkj} such that $E(\eta_{tkj}^2) = 0$ the corresponding regression coefficients are fixed over time. Notice also that since $\underset{\sim}{\beta}_{tk} = \underset{\sim}{\beta}_{t-1,k} + \underset{\sim}{\eta}_{tk}$, the predictor of $\underset{\sim}{\beta}_{tk}$ at time $(t - 1)$ is the same as the predictor $\hat{\underset{\sim}{\beta}}_{t-1,k}$ of $\underset{\sim}{\beta}_{t-1,k}$.

Using the Kalman filter equations presented in section 3, it is shown in the Appendix that the estimator $\hat{\Theta}_{tk}$ of the small area mean Θ_{tk} (equation 2.2) can be structured in this case in the following form

$$\hat{\Theta}_{tk} = \underset{\sim}{x}_{tk}'\hat{\underset{\sim}{\beta}}_{t-1,k} + \left(1 - \frac{\sigma_k^2}{n_{tk}\nu_k^2}\right)\left(\bar{Y}_{tk} - \underset{\sim}{x}_{tk}'\hat{\underset{\sim}{\beta}}_{t-1,k}\right) + \frac{\sigma_k^2}{n_{tk}\nu_k^2}\sum_{\substack{m=1\\ m\neq k}}^{K} \gamma_{km}\left(\bar{Y}_{tm} - \underset{\sim}{x}_{tm}'\hat{\underset{\sim}{\beta}}_{t-1,m}\right) \tag{2.12}$$

where the coefficients $\{\gamma_{km}\}$ are the partial regression coefficients in the regression of $e_{tk} = (\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$ against the prediction errors $\{e_{tm} = (\bar{Y}_{tm} - \underset{\sim}{x}'_{tm}\hat{\underset{\sim}{\beta}}_{t-1,m})\}$ obtained in the other areas and v_k^2 is the residual (unexplained) variance in the regression.

The estimator $\hat{\Theta}_{tk}$ is composed of three components: the "synthetic" estimator, $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$, where $\hat{\underset{\sim}{\beta}}_{t-1,k}$ is the optimal predictor of $\underset{\sim}{\beta}_{tk}$ based on all the observations up to and including time $t-1$, the "correction factor" $(\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$ based on the prediction error in area k, and an "adjustment factor" based on the prediction errors observed for the other areas. The first two components correspond to the components of the classical small area estimators discussed in the introduction. Notice that the smaller the sample size n_{tk}, the smaller is the weight assigned to the current sample mean $\bar{Y}_{tk}$ in the estimation of Θ_{tk} and the larger is the weight assigned to the time series predictor $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$. The third component in the right hand side of (2.12) represents the information borrowed from neighbouring areas. The weight assigned to this component depends on the magnitude of the correlations ρ_j between the corresponding error terms $\{\eta_{tkj}\}$ in the models holding for the regression coefficients (equation 2.11). Obviously, when the regressions in the various areas are independent so that $\rho_j = 0$ for all j and hence $\gamma_{km} = 0$ for all m, the third component vanishes and the predictor $\hat{\Theta}_{tk}$ reduces to a weighted average of the current mean $\bar{Y}_{tk}$ and the time series predictor $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$.

3. MODEL ESTIMATION AND INITIALIZATION USING THE KALMAN FILTER

3.1 Estimation of the Regression Coefficients by Means of the Kalman Filter

In this section we present the Kalman filter equations for the updating and smoothing of the state vectors $\underset{\sim}{\alpha}_t$ defined by the equations (2.7) and (2.8) (the area regression coefficients in our case). We assume that the V-C matrices Σ_t and Λ are known. Estimation of these matrices is considered in section 3.2. 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 restrict the discussion to aspects most germane to the small area estimation problem.

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 $(t-1)$, $P_{t|t-1} = TP_{t-1}T' + G\Lambda G'$ is the V-C matrix of the prediction errors $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$. (Follows straightforwardly from 2.8).

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

$$\begin{aligned} \hat{\underset{\sim}{\alpha}}_t &= \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1}Z'_tF_t^{-1}(\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1}) \\ P_t &= (I - P_{t|t-1}Z'_tF_t^{-1}Z_t)P_{t|t-1} \end{aligned} \tag{3.1}$$

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 $F_t = (Z_tP_{t|t-1}Z'_t + \Sigma_t)$.

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

$$\hat{\underset{\sim}{\alpha}}_{t|t^*} = \hat{\underset{\sim}{\alpha}}_t + P_t T' P_{t+1|t}^{-1}(\hat{\underset{\sim}{\alpha}}_{t+1|t^*} - T\hat{\underset{\sim}{\alpha}}_t)$$
$$P_{t|t^*} = P_t + P_t T' P_{t+1|t}^{-1}(P_{t+1|t^*} - P_{t+1|t})P_{t+1|t}^{-1} T P_t; \; t = 2, 3, \ldots, t^* \tag{3.2}$$

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

Estimators of the small area means or aggregates of the means are obtained from the filtered (or smoothed) estimators of $\underset{\sim}{\alpha}_t$ in a straightforward manner using the relationship $\hat{\Theta}_{tk} = \bar{\underset{\sim}{X}}_{tk}\hat{\underset{\sim}{\beta}}_{tk} = \bar{\underset{\sim}{Z}}'_{tk}\hat{\underset{\sim}{\alpha}}_{tk} = \bar{\underset{\sim}{Z}}'_{tk}A_{tk}\hat{\underset{\sim}{\alpha}}_t$ where $\bar{\underset{\sim}{Z}}'_{tk} = (1, 0, \bar{X}_{tk1}, 0, \ldots, \bar{X}_{tkp}, 0)$ and A_{tk} is the appropriate indicator matrix. Hence, if $\Theta_t^w = \sum_{k=1}^K w_k\Theta_{tk}$, then $\hat{\Theta}_t^w = \sum_{k=1}^K w_k\bar{\underset{\sim}{Z}}'_{tk}A_{tk}\hat{\underset{\sim}{\alpha}}_t = \underset{\sim}{a}'_{tw}\hat{\underset{\sim}{\alpha}}_t$, say. For given V-C matrices Σ_t and Λ, the MSE's of the estimation errors are obtained as

$$E(\hat{\Theta}_{tk} - \Theta_{tk})^2 = \bar{\underset{\sim}{Z}}'_{tk}A_{tk}P_tA'_{tk}\bar{\underset{\sim}{Z}}_{tk} \quad \text{and} \quad E(\hat{\Theta}_{tk}^w - \Theta_{tk}^w) = \underset{\sim}{a}'_{tw}P_t\underset{\sim}{a}_{tw}. \tag{3.3}$$

Notice that the MSE's in (3.3) are with respect to the joint distribution of the observations $\{\underset{\sim}{Y}_{tk}\}$ and the vectors of coefficients $\{\underset{\sim}{\beta}_{tk}\}$ so that they represent average MSE's over the possible realizations of the area means.

3.2 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 simple estimation procedures which can be used for these purposes.

Assuming a normal distribution for the residual terms $\underset{\sim}{\epsilon}_t$ and η_t of equations (2.7) and (2.8), the log likelihood function of the vectors $\underset{\sim}{Y}_{m+1}, \ldots, \underset{\sim}{Y}_{t^*}$, conditional on the first m vectors $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_m$, can be formulated as

$$L(\underset{\sim}{\lambda}) = \text{constant} - \frac{1}{2}\sum_{t=m+1}^{t^*}(\log|F_t| + \underset{\sim}{e}'_tF_t^{-1}\underset{\sim}{e}_t) \tag{3.4}$$

where $\underset{\sim}{\lambda}$ contains the unknown model variances and covariances written in a vector form. The scalar m defines the number of time periods needed to construct initial values for the Kalman filter. (For the random walk model considered in section 2.2, $m = 1$, provided that sufficient data are available in every area to allow the computation of the OLS estimators of the vectors of coefficients). The expression in (3.4) follows from 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 can be obtained by application of the Kalman filter equations (3.1).

The computation of the likelihood function requires the initialization of the Kalman filter which can be carried out most conveniently by application of the approach proposed by Harvey and Phillips (1979). By this approach, the nonstationary components of the state vector are initialized with very large error variances which corresponds to postulating a noninformative prior distribution so that the corresponding state estimates can conveniently be taken as zeroes. (For the random walk model, initializing with a noninformative prior yields the OLS estimators after one time period, see Meinhold and Singpurwalla 1983, for a Bayesian formulation of the Kalman filter). The stationary components of the state vector are initialized by the corresponding unconditional means and variances which may be part of the unknown parameters defining the arguments of the likelihood function.

Maximization of the likelihood function (3.4) can be implemented using the method of scoring with a variable step length. In particular, let $\underset{\sim}{\lambda}_{(o)}$ define initial estimates of the unknown elements in $\underset{\sim}{\lambda}$. Then the method of scoring consists of solving iteratively the set of equations

$$\underset{\sim}{\lambda}_{(i)} = \underset{\sim}{\lambda}_{(i-1)} + r_i\{I[\underset{\sim}{\lambda}_{(i-1)}]\}^{-1} g[\underset{\sim}{\lambda}_{(i-1)}] \tag{3.5}$$

where $\underset{\sim}{\lambda}_{(i-1)}$ is the estimator of $\underset{\sim}{\lambda}$ as obtained in the $(i-1)$-th iteration, $I[\underset{\sim}{\lambda}_{(i-1)}]$ 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 a 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 can be determined by a grid search procedure in the region $[0,1]$. The formulae for the k-th element of the gradient vector and the $k\ell$-th element of the information matrix are given in Watson and Engle (1983).

Having estimated the model variances and covariances, these estimates can be substituted for the true parameters in the Kalman filter equations (3.1) - (3.2) to yield the estimators of the regression coefficients and the V-C matrices and hence the small area estimators and their variances (see equation 3.3). Notice however that the estimated V-C matrices ignore the 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 using first order Taylor approximations. Hamilton (1986) proposes a Monte Carlo procedure which consists of sampling from a multivariate normal distribution with mean given by the maximum likelihood estimator of the vector $\underset{\sim}{\lambda}$ and V-C matrix defined by the inverse of the information matrix, and estimating the state vectors for each random realization of the parameter values. This procedure is more flexible in terms of the assumptions involved and provides further insight into the sensitivity of the Kalman filter estimators to errors in the variance and covariance estimators. However, it is computationally more intensive.

4. MODIFICATIONS TO PROTECT AGAINST MODEL BREAKDOWNS

4.1 Description of the Problem and Proposed Modifications

The use of a model for small area estimation seems inevitable in view of the small sample sizes within the areas. However it raises the question of how to protect against model breakdowns. Testing the model every time that new data becomes available is often not practical, requiring instead the development of a "built-in mechanism" to ensure the robustness of the estimators when the model fails to hold.

One possibility is 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 fitted values under the model. More precisely, we propose to augment the model equation (2.1) by linear constraints of the form

$$\sum_k W_{tk}^{(\ell)} \sum_i Y_{tki} = \sum_k W_{tk}^{(i)} \sum_i \underset{\sim}{x}'_{tki} \underset{\sim}{\beta}_{tk} \quad \ell = 1, 2, \ldots, L(t),\ t = 1, \ldots, t^* \tag{4.1}$$

where the coefficients $W_{tk}^{(\ell)}$ are fixed, standardized weights such that $\sum_k n_{tk} W_{tk}^{(\ell)} = 1$. An example for such a constraint would be the equation

$$\sum_{k=1}^{K} N_{tk}\hat{M}_{tk} \Big/ \sum_{k=1}^{K} N_{tk} = \sum_{k=1}^{K} N_{tk}(\bar{\underset{\sim}{x}}'_{tk}\underset{\sim}{\beta}_{tk}) \Big/ \sum_{k=1}^{K} N_{tk} \tag{4.2}$$

where $\hat{M}_{tk}$ is the direct, survey estimator in area k. For $\bar{\underset{\sim}{x}}_{tk} = \bar{\underset{\sim}{X}}_{tk}$, the equation (4.2) guarantees that the model dependent predictor of the aggregate population mean coincides with the corresponding survey estimator. Such a constraint can be justified by arguing that the survey estimators, although not reliable enough for estimating the small area means due to the small sample sizes, can be trusted when being combined for estimating the aggregate mean. Notice that "adding up" constraints are ordinarily imposed on statistical agencies anyway. Battese, Harter and Fuller (1988) and Pfeffermann and Barnard (1991) use a similar constraint for analysing cross-sectional surveys. Often, the small areas can be grouped into broader groups, with sufficient data in each of the groups to justify the use of the survey estimators for estimating the corresponding group means. In this case, one can impose several constraints of the form (4.2) where the summation is now over the areas belonging to the same group. Notice in this respect that in view of the correlations between the regression coefficients operating in the various areas, a constraint applied to a sub-set of the areas will modify the regression estimates in all the areas. We illustrate this property in the empirical study.

It is important to emphasize that the set of constraints in (4.1) does not represent external information about possible values of the regression coefficients. Rather, it serves as a "control system" to guarantee that the model estimators adjust themselves more rapidly to possible changes in the behavior of the regression coefficients. As a result, 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 even without imposing them explicitly. As mentioned above, it is possible 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 needed and do not interfere with the random fluctuation of the raw data.

4.2 Inference Incorporating the Linear Constraints

In Section 4.1 we proposed to amend the model equations (2.1) by imposing the set of constraints (4.1) thereby ensuring the robustness of the regression estimators against sudden drifts in the values of the coefficients.

Computationally, this can be implemented most conveniently by augmenting the vectors $\underset{\sim}{Y}_t$ of equation (2.7) by the scalars $\sum_k W_{tk}^{(t)} \sum_i Y_{tki}$, augmenting the matrices Z_t by the corresponding row vectors $(W_{t1}^{(t)} 1'_{nt1} Z_{t1}, \ldots, W_{tK}^{(t)} 1'_{ntK} Z_{tK})$ and setting the respective variances of the residual terms to zero. The augmented set of equations, together with (2.8), form a pseudo state-space model which could be estimated using the Kalman filter equations (3.1). Notice that the pseudo V-C matrix $\Sigma_t^{(P)}$ of the augmented residual vector is no longer positive definite (the last $L(t)$ rows and columns of $\Sigma_t^{(P)}$ consist of zeroes) but this does not cause 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 appearing in the left hand side of (4.1). In order to deal with this problem, we propose to amend the formula for the updating of the V-C matrix P_t (equation 3.1) so that the variances and covariances of the aggregate means 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 $[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t]$. The matrix $\Sigma_t^{(A)}$ is of order $[n_t + L(t)]$ with Σ_t in the first n_t rows and columns and the variances and covariances of the means $\Sigma_k W_{tk}^{(l)} \Sigma_i Y_{tki}$ 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

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + P_{t|t-1}^{(A)} Z_t^{(A)\prime}\left(F_t^{(P)}\right)^{-1}\left[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}\right] \tag{4.3}$$

where $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)}T' + G\Lambda G')$ and $F_t^{(P)} = Z_t^{(A)} P_{t|t-1}^{(A)} Z_t^{(A)\prime} + \Sigma_t^{(P)}$ (Compare with 3.1). It is shown in the Appendix 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

$$P_t^{(A)} = \left[I - K_t^{(P)} Z_t^{(A)}\right] P_{t|t-1}^{(A)} + K_t^{(P)}\left[\Sigma_t^{(A)} - \Sigma_t^{(P)}\right] K_t^{(P)}, \tag{4.4}$$

where $K_t^{(P)} = P_{t|t-1}^{(A)} Z_t^{(A)\prime} (F_t^{(P)})^{-1}$ is the pseudo Kalman gain. The first expression on the right hand side of (4.4) corresponds to the usual updating formula of the Kalman filter (compare with 3.1)). The second expression is a correction factor which accounts for the actual variances and covariances of the means $\Sigma_k W_{tk}^{(l)} \Sigma_i Y_{tki}$, not taken into account in the first expression.

The amended Kalman filter defined by the equations (4.3) and (4.4) produces robust predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ instead of the optimal, model dependent predictors, $\hat{\underset{\sim}{\alpha}}_t$ but otherwise 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 hence for the estimation of the small area means, 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 (4.4) could be incorrect (depending on the particular model failures) but the predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ will nonetheless satisfy the linear constraints (4.1). The smoothing equations (3.2) can likewise be modified to satisfy the linear constraints.

5. EMPIRICAL RESULTS

5.1 Description of the Data and Model Fitted

In order to illustrate the important features of the class of models defined in Section 2, we fitted such a model to home sale prices in Jerusalem. The sale prices are recorded on a monthly basis and are routinely used by the Central Bureau of Statistics in Israel for the computation of monthly housing price indexes (HPI) adjusted for changes in quality. The HPI is computed separately for each city or group of cities and for each house size defined by the number of rooms, ranging from 1 to 5. The number of transactions carried out each month is very small in many of these cells and for 1 room apartments it occasionally happens that there are no transactions. The mean and standard deviation (S.D.) of the monthly number of transactions carried out during the period July 1987 - November 1989 are listed below.

Size	1	2	3	4	5
Mean	2.7	29.0	101.9	39.7	5.6
S.D.	2.6	12.9	50.4	18.8	3.5

The need to adjust for changes in quality results from the fact that the transactions performed are not under control, giving rise to large differences in quality from one month to the other particularly in the small cells. The following quality measure variables (QMV) are recorded for every transaction: $\tilde{X}^{(1)}$ - the apartment floor area, $\tilde{X}^{(2)}$ - the age of the apartment, $X^{(3)}$, $X^{(4)}$ - dummy variables defining districts within the city.

The problems involved in the computation of the HPI and the method used in Israel are discussed at length in a recent article by Pfeffermann, Burck and Ben-Tuvia (1989). The following model was proposed by the authors as an alternative to the model in current use. The triple index "tki" defines the i-th transaction of size k in month t with Y_{tki} standing for the log of the sale price and $X_{tki}^{(j)} = \log(\tilde{X}_{tki}^{(j)})$, $j = 1, 2$.

$$Y_{tki} = \beta_{tk0} + \beta_{tk1}X_{tki}^{(1)} + \beta_{tk2}X_{tki}^{(2)} + \beta_{tk3}X_{tki}^{(3)} + \beta_{tk4}X_{tki}^{(4)} + \epsilon_{tki} \tag{5.1}$$

$$\begin{aligned} \beta_{tk0} &= \beta_{t-1,k0} + \beta_{k0} + \eta_{tk0} \\ \beta_{tkj} &= \beta_{t-1,kj} + \eta_{tkj},\ j = 1, \ldots, 4, \end{aligned} \tag{5.2}$$

with the error terms ϵ_{tki} and η_{tkj} satisfying the assumptions (2.1), (2.4) and (2.5). Notice that the model assumed for the intercept term is the local approximation to a linear trend defined under case (d) of Section (2.1). The model assumed for the other coefficients is the random walk model defined under case (b).

The regression defined by (5.1) forms the basis for the construction of an HPI adjusted for changes in quality. By fixing the values of the QMV's at their average population values which are constant over time, (the values of these variables are adjusted approximately every five years), average sale prices can be computed using (5.1) and these averages are comparable between months since they refer to homes of similar qualities.

Pfeffermann, Burck and Ben-Tuvia discuss the considerations in selecting the model defined by (5.2) for the regression coefficients. They show empirical results which validate the fitness of the model. However, the results of that study were obtained by fitting the model to each cell separately, that is, without accounting for the cross-sectional relationships of the regression coefficients. This aspect of the model is explored in the present study. Another major purpose of the empirical study is to illustrate the performance of the modifications proposed in Section 4 to protect against model breakdowns.

5.2 Estimation of the Model

The model defined by (5.1) and (5.2) can be put in a state-space form similar to (2.7) and (2.8). In fact, the vectors $\underset{\sim}{\alpha}_t$ and the matrices Z_t, T and G assume, in this case, simple structures, since for $j = 1, \ldots, 4$, $\beta_{kj} \equiv 0$ (see case (b) of Section 2.1). Thus, $\underset{\sim}{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \ldots, \beta_{tk4})$, $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, \underset{\sim}{X}_{tk}^{(1)}, \ldots, \underset{\sim}{X}_{tk}^{(4)}]$, $\tilde{T} = [\underset{\sim}{e}_1, \underset{\sim}{e}_1 + \underset{\sim}{e}_2, \underset{\sim}{e}_3, \ldots, \underset{\sim}{e}_6]$, a 6×6 matrix with $\underset{\sim}{e}_j$ having a one in position j and zeroes elsewhere and $\tilde{G} = [\underset{\sim}{e}_1, \underset{\sim}{e}_3, \ldots, \underset{\sim}{e}_6]$ which is 6×5. The matrix Δ is defined as in (2.5). The vector $\underset{\sim}{\alpha}_t$ and the matrices Z_t, T, G and Λ are obtained from the vectors $\{\underset{\sim}{\alpha}_{tk}\}$ and the matrices $\{Z_{tk}\}$, $\tilde{T}$, $\tilde{G}$ and Δ in the same way as in (2.7) and (2.8).

Having set the model in a state-space form we next attempted to estimate the unknown variances and covariances using the method of scoring algorithm described in Section 3.2. As it turned out, however, the computer time needed for convergence was way beyond the capacity of the IBM 1481 mainframe used for this study. Notice that the number of unknown parameters of the combined state-space model is $\dim(\underset{\sim}{\lambda}) = 25$ whereas the dimension of the

state vectors and hence the dimension of the corresponding V-C matrices is $\dim(\underset{\sim}{\alpha}_t) = 30$. The total number of observations per month ranges from 55 to 353. The computer program written for this study uses numerical derivatives so that each iteration of the method of scoring requires a separate sweep through all the data with each sweep involving $[\dim(\underset{\sim}{\lambda}) + 1]$ computations of the state vector $\hat{\underset{\sim}{\alpha}}_t$ and the V-C matrix P_t (equation 3.1) at each point in time. These computations are needed in order to evaluate the log likelihood functions and hence the corresponding derivatives. It is clear therefore that the computational costs increase with the length of the series, the number of observations, the size of the state vector and the number of unknown parameters.

In order to deal with this problem we estimated the variance σ_k^2 (equation 2.1) and the matrix Δ (equation 2.5) separately for each of the five apartment sizes using the time series of observations corresponding to each size and then estimated the correlations ρ_j (equation 2.6) by a crude, grid search procedure. We found that setting $\rho_j = ½$ for every j gives satisfactory results both in terms of the behaviour of the innovations (the one step ahead prediction errors) and in terms of the smoothness of the regression coefficients corresponding to apartments of size one and five where the monthly sample sizes are very small. Notice that by estimating the variances and covariances defining the time series relationships of the regression coefficients separately for each size, one is more flexible in terms of the model assumptions although there is some loss of efficiency if the variances and covariances are indeed the same across the different sizes.

5.3 Results

Pfeffermann, Burck and Ben-Tuvia (1989) illustrate the adequacy of the time series models fitted to the various apartment sizes. As mentioned earlier, our purpose in this study is to compare the results obtained with and without the accounting for the cross-sectional correlations and to illustrate the performance of the modifications (4.1) in protecting against model breakdowns.

In order to sharpen the comparisons as much as possible, we deliberately inflated the Y-values by 5 percent in each of the following four months: October 1987, November 1988, January 1989 and May 1989. Thus all the Y-values of all the apartment sizes corresponding to the months October 1987 - October 1988 were inflated by 5 percent, the Y-values corresponding to November 1988 - December 1988 were inflated by 10.25 percent (5 percent on top of the previous 5 percent) and so forth. These kinds of model breakdowns (although obviously not in such magnitudes) may result from intentional devaluations of the currency and are of main concern when modeling sale prices. See Pfeffermann, Burck and Ben-Tuvia for further discussion. Similar model breakdowns may occur, for example, with series of unemployment rates in periods of abrupt economic recessions.

Table 1 shows the average mean squared errors (AMSE) of the model residuals $\hat{\epsilon}_{tki} = (Y_{tki} - \hat{\beta}_{tko} - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{tkj})$ and the model innovations $e_{tki} = [Y_{tki} - (\hat{\beta}_{t-1,k0} + \beta_{k0}) - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{t-1,kj}]$ (see equations 5.1 and 5.2), separately for each of the five apartment sizes. The AMSE's were computed as $\text{AMSE}_k(\epsilon) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} \hat{\epsilon}_{tki}^2)$; $\text{AMSE}_k(e) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} e_{tki}^2)$ where $t = 1, \ldots, N$ indexes the months of July 1987 - November 1989. We distinguish between four different estimators of the regression coefficients as defined by whether the model accounts for the cross-sectional correlations ($\rho_j \equiv ½$), ($\rho_j \equiv 0$) and by whether or not the estimators are modified to protect against the model breakdowns (abbreviated as "Rob. Inc." and "No Rob." in the table). The modifications were carried out by augmenting the observation equation of each month by three linear constraints of the form 4.2. These constraints forced the aggregate means of the fitted values in each of the three

Table 1

Average Mean Squared Errors of Residuals and Innovations With and Without the Accounting for Cross-sectional Correlations and the Inclusion of the Robustness Modifications, by Size

Apt. Size	Mean Squared Errors of Innovations				Mean Squared Errors of Residuals			
	$\rho \equiv ½$		$\rho \equiv 0$		$\rho \equiv ½$		$\rho \equiv 0$	
	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.
1	.141	.134	.176	.218	.021	.027	.056	.092
2	.070	.090	.084	.123	.021	.039	.023	.070
3	.065	.090	.070	.197	.017	.042	.019	.143
4	.067	.123	.072	.198	.019	.066	.021	.141
5	.067	.114	.077	.193	.023	.033	.065	.106

districts to coincide with the corresponding means of the observed values. When incorporating the constraints, the model was fitted using the amended Kalman filter as defined by the equations (4.3) and (4.4).

In order to illustrate the performance of the four sets of regression estimators in the various months and in particular, in and around the months where we inflated the data, we plotted the monthly MSE's of the innovations and residuals as obtained for 3 and 5 room apartments. The plots are shown in Figures 1 to 4. Notice that the values of Table 1 for 3 and 5 room apartments are correspondingly the averages of the values shown in the four figures.

The main conclusions from the table and the graphs are as follows:

Accounting for the cross-sectional correlations and including the linear constraints to protect against the model breakdowns yields better results than in the other cases considered. This outcome is most prominent in the cells of 1 and 5 room apartments where the sample sizes in each month are very small. In the other three cells, there are only small differences between the case ($\rho \equiv ½$, Rob. Inc.) and the case ($\rho \equiv 0$, Rob. Inc.) which could be expected since as the number of observations in each month increases, there is less borrowing of information from neighbouring cells (small areas in the more general context). The situation is different, however, when the linear constraints are removed. Accounting for the cross-sectional correlations yields in this case much better results than when not accounting for them and this is true for all the apartment sizes. Thus, by borrowing information from one cell to the other, the estimators of the regression coefficients adapt themselves much more rapidly to the sudden drifts in the data as seen also more directly in the figures [The four peaks in each graph are in the months where the data were inflated and as can be seen, the graphs corresponding to the case ($\rho \equiv ½$, No Rob.) return to their normal level of the months before the inflation much faster than the graphs representing the case ($\rho \equiv 0$, No Rob.)

Another interesting comparison is between the case where the linear constraints are included and the case where they are not. Clearly, the inclusion of the constraints improves the results substantially when accounting for the serial correlations and the improvements are even more prominent when the serial correlations are set to zero. It is interesting to compare in this context the figures exhibiting the monthly MSE's of the innovations with the figures exhibiting the monthly MSE's of the residuals. In the four months where we inflated the data the MSE's of the innovations are high which is obvious since the innovations are the differences between the observations and their predictors from previous months. Still, when the linear constraints are included, the MSE's return to their normal level right after the months of inflation. As

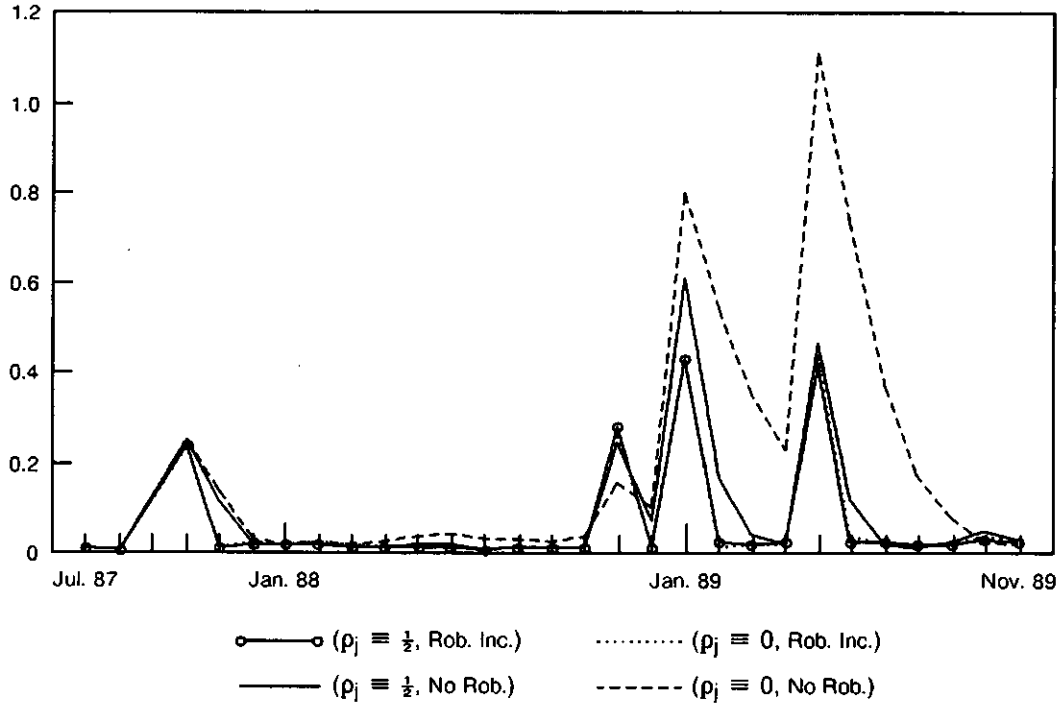


Figure 1 Monthly Mean Squared Errors of Innovations, 3 Room Apartments

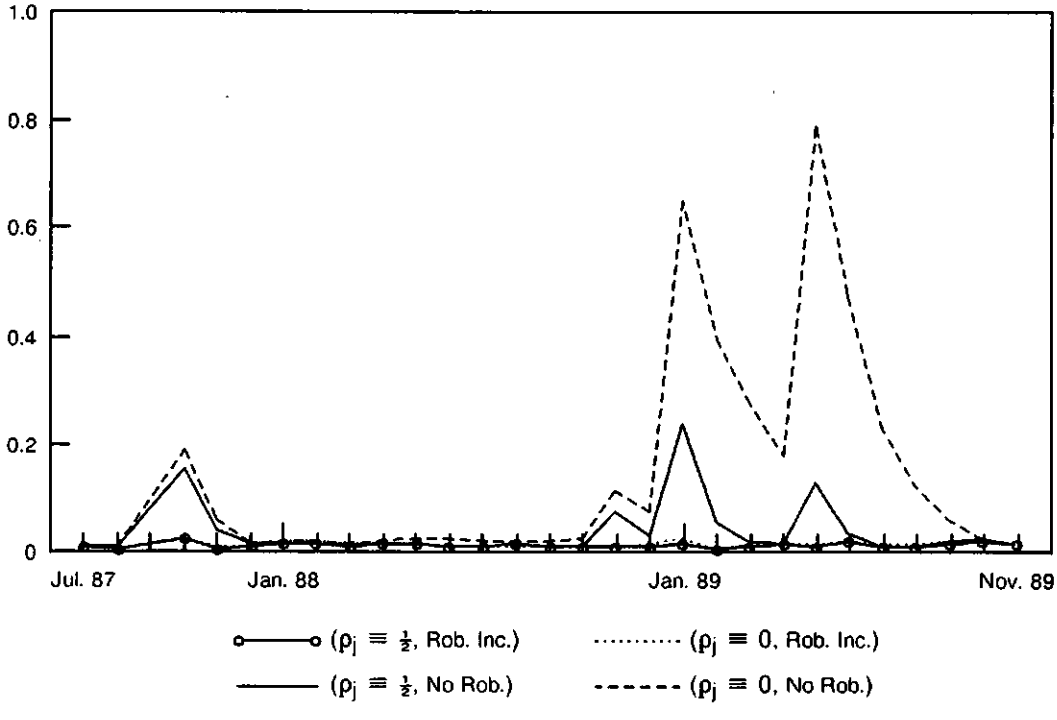


Figure 2 Monthly Mean Squared Errors of Residuals, 3 Room Apartments

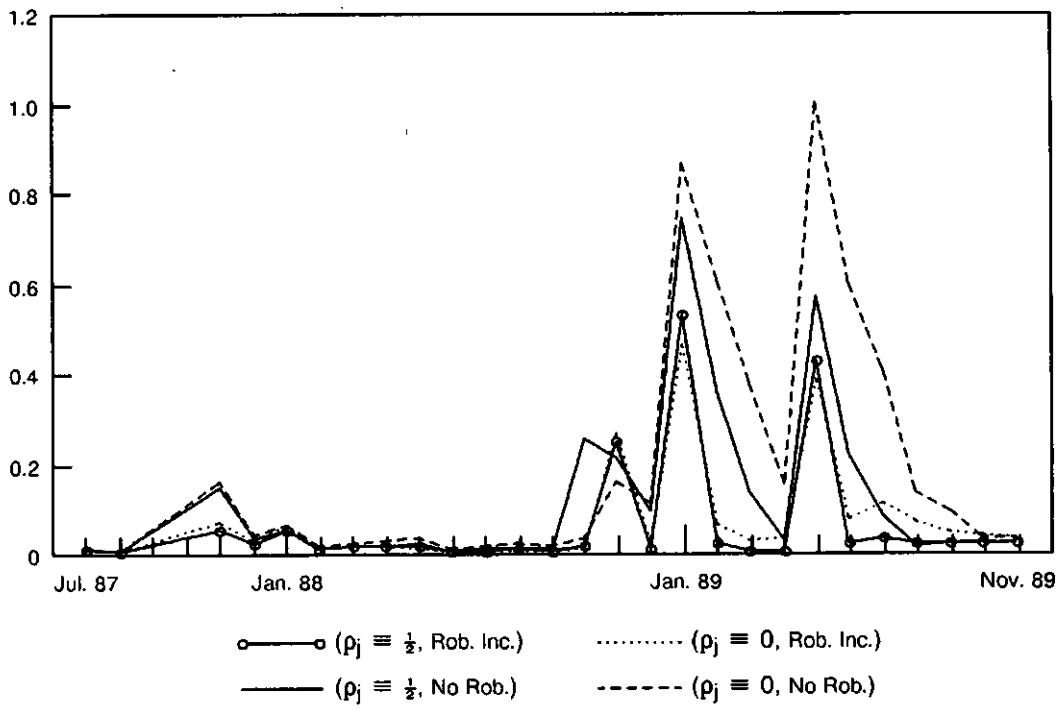


Figure 3 Monthly Mean Squared Errors of Innovations, 5 Room Apartments

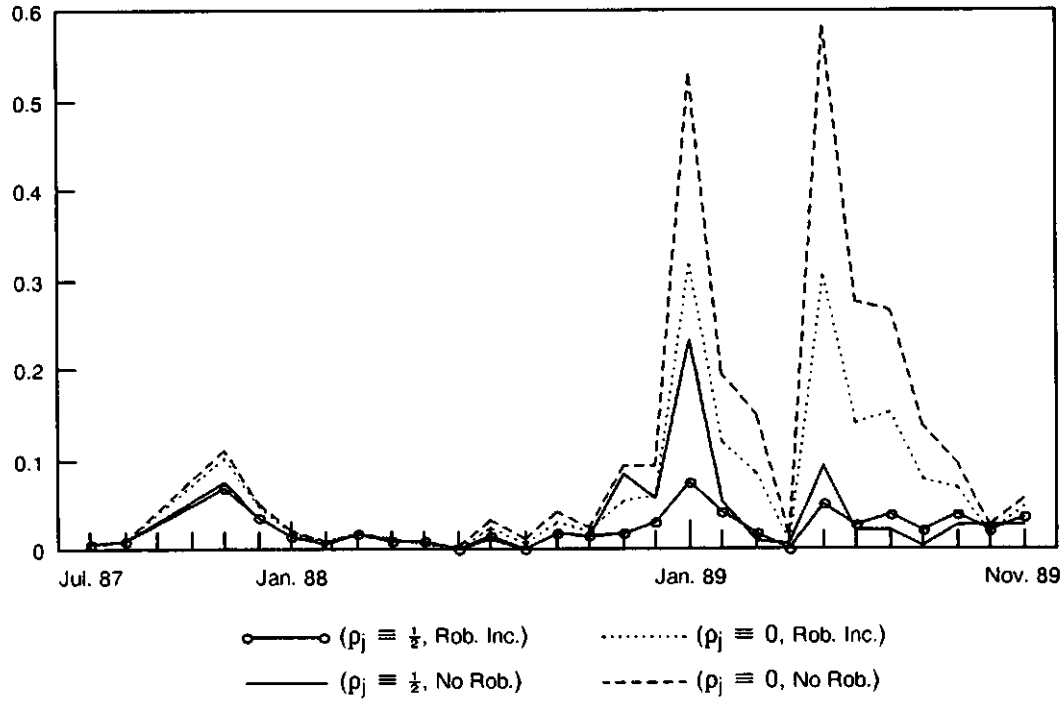


Figure 4 Monthly Mean Squared Errors of Residuals, 5 Room Apartments

for the residuals, once the linear constraints are included, there is practically no increase in the MSE values in the months of inflation in the case of 3 room apartments and, when accounting for the serial correlations, only a slight increase in the case of 5 room apartments. However, when ignoring the serial correlations, the residual MSE's for 5 room apartments are much larger in the months of inflation than in the other months even when imposing the constraints. This outcome has a simple explanation. The linear constraints are imposed on the aggregate means of the fitted values in each district but since the number of observations in 5 room apartments is a small fraction of the total number of observations, the constraints alone have a relatively small effect on the estimated regression coefficients in this cell. On the other hand, the constraints have a large effect on the estimated coefficients in the other cells so that when accounting for the cross-sectional correlations, the estimators corresponding to 5 room apartments are also modified since they are correlated with the other coefficients.

The way by which the linear constraints protect against sudden drifts in the data is illuminated in Figure 5 where we plotted the monthly intercept estimates for 3 room apartments.

As can be seen, with the linear constraints included, the intercept adapts itself to the new level of the data in the same month that the inflation occurs. Without the inclusion of the constraints, the adaption to the new level of the data takes several months. The plot of the monthly intercept estimates of 5 room apartments does not have this nice pattern since with the small sample sizes observed each month, the effect of the inflation is to alter also the other regression coefficients.

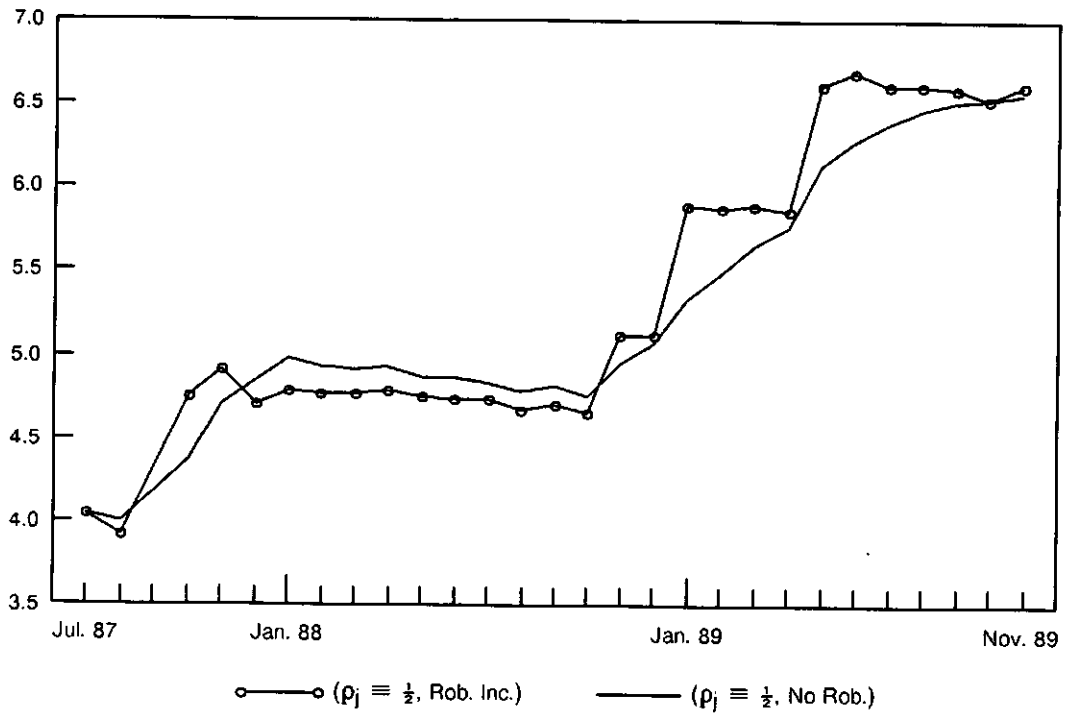


Figure 5 Monthly Estimates of Intercept, 3 Room Apartments

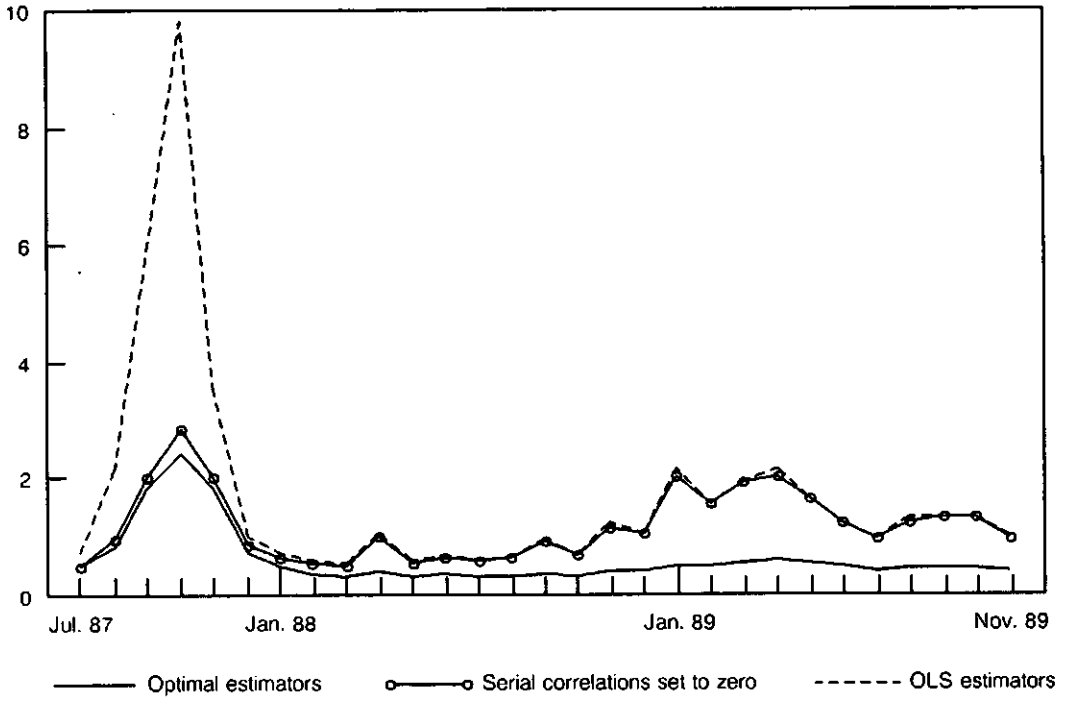


Figure 6 Variances of Estimators of Cell Means ($\times 10^4$), 3 Room Apartments

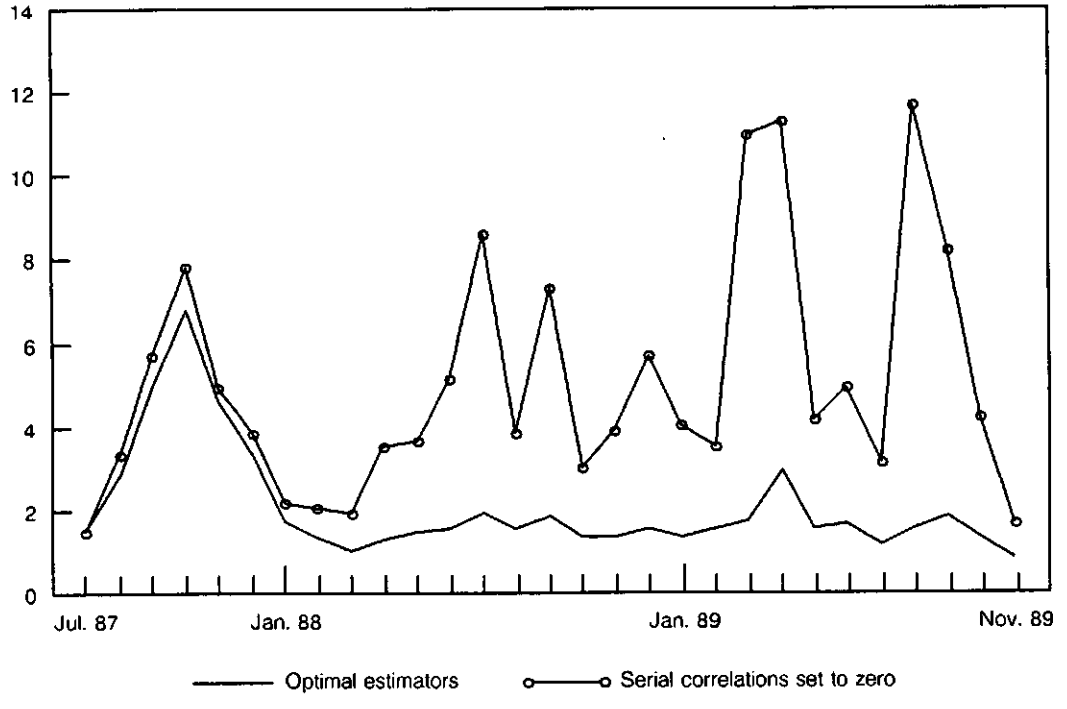


Figure 7 Variances of Estimators of Cell Means ($\times 10^4$), 5 Room Apartments

Our discussion so far centered on the empirical distribution of the model residuals and innovations. A major application of small area estimation is the prediction of the small area means (equation 2.2). Clearly, when a model yields residuals with well behaved properties it can also be expected to yield good estimators for the population means. Nevertheless, it is interesting to compare the theoretical variances of the small area means estimators as obtained with and without the accounting for the cross-sectional correlations, under the model which accounts for these correlations with $\rho_j \equiv \frac{1}{2}$. This comparison permits the assessment of the loss in efficiency when the serial correlations are ignored.

Figures 6 and 7 show the monthly variances of the cell mean estimators as obtained for 3 and 5 room apartments. (The variances have been multiplied by 10^4.) The figure for 3 room apartments also contains the variances of the ordinary least squares (OLS) estimators of the population means, that is, the variances of the estimators when estimating the regression coefficients in each month by OLS. These estimators are not operational in the case of 5 room apartments because of the very small monthly sample sizes.

The important conclusion drawn from the two figures is that by accounting for the cross-sectional correlations the variances of the resulting estimators can be reduced quite substantially, depending on the sample sizes. This is obviously the case in the case of 5 room apartments but is also true for 3 room apartments despite the fact that the sample sizes in these cells are relatively very large. The large sample sizes ordinarily obtained for 3 room apartments make the OLS estimators quite comparable to the estimators obtained when ignoring the cross-correlations in the estimation of the population means. Notice however the big gap between the variance of the OLS estimator and the variance of the other two estimators in October 1987. In this month there were only 10 observations of 3 room apartments and it is here where the use of the past data has its main impact even when ignoring the cross-sectional correlations. (The number of observations for 3 room apartments in November 1987 is 28; in all the other months there are at least 46 observations.)

Another important outcome arising from the two figures is the much greater stability of the variances of the optimal estimators under the model as compared to the variances of the estimators which ignore the cross-sectional correlations. Notice in this respect that the differences in the variances from one month to the other depend not only on the sample sizes in each month but also on the values of the explanatory variables (the design matrix) and the amount of past data observed. Still, it is the sample sizes which mostly explains the differences in the variances of the estimators particularly towards the end of the series.

ACKNOWLEDGEMENT

This article was written while the first author was on sabbatical leave at Statistics Canada under its Research Fellowship program. The authors would like to thank a referee for helpful comments.

APPENDIX

a) Derivation of Equation (2.12)

When $\underset{\sim}{x}_{tki} = \underset{\sim}{x}_{tk}$, $\hat{\Theta}_{tk} = \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{tk} = \underset{\sim}{z}'_{tk}\hat{\underset{\sim}{\alpha}}_{tk}$ so that $\hat{\underset{\sim}{\Theta}}_t = (\hat{\Theta}_{t1}, \ldots, \hat{\Theta}_{tK})' = Z_t\hat{\underset{\sim}{\alpha}}_t$. Also, for the random walk model the matrix T is the identity matrix and by equation (3.1)

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

$$\left(I - \Sigma_tF_t^{-1}\right)\underset{\sim}{Y}_t + \Sigma_tF_t^{-1}Z_t\hat{\underset{\sim}{\alpha}}_{t-1} \tag{A1}$$

since $F_t = (Z_tP_{t|t-1}Z'_t + \Sigma_t)$. Suppose for convenience that $k = 1$ and define

$$F_t = \begin{bmatrix} f_{11}, \underset{\sim}{f}'_1 \\ \underset{\sim}{f}_1, F_{22} \end{bmatrix} \quad \text{and} \quad H_t = F_t^{-1} = \begin{bmatrix} h_{11}, \underset{\sim}{h}'_1 \\ \underset{\sim}{h}_1, H_{22} \end{bmatrix} \quad \text{were } f_{11} \text{ and } h_{11}$$

are scalars, $\underset{\sim}{f}'_1$ and $\underset{\sim}{h}'_1$ are $[1 \times (K-1)]$ and F_{22} and H_{22} are $[(K-1) \times (K-1)]$. Using this notation, it follows from (A1) that

$$\hat{\Theta}_{t1} = \left(1 - \frac{\sigma_1^2}{n_{t1}}h_{11}\right)\bar{Y}_{t1} + \frac{\sigma_1^2}{n_{t1}}h_{11}\left(\underset{\sim}{x}'_{t1}\hat{\underset{\sim}{\beta}}_{t-1,1}\right) - \frac{\sigma_1^2}{n_{t1}}\sum_{k=2}^{K} h_{11}\frac{h_{1k}}{h_{11}}\bar{e}_{tk}. \tag{A2}$$

Let $\underset{\sim}{\gamma}'_1 = (\gamma_{12}, \ldots, \gamma_{1K}) = \underset{\sim}{f}'_1F_{22}^{-1}$ defines the partial regression coefficients in the regression of $\bar{e}_{t1}$ on $(\bar{e}_{t2}, \ldots, \bar{e}_{tK})$ and $v_1^2 = (f_{11} - \underset{\sim}{f}'_1F_{22}^{-1}\underset{\sim}{f}_1)$ define the residual variance in the regression.

Equation (2.12) follows directly from (A2) since

$$\underset{\sim}{f}'_1F_{22}^{-1} = -\frac{1}{h_{11}}\underset{\sim}{h}'_1; \quad \left(f_{11} - \underset{\sim}{f}'_1F_{22}^{-1}\underset{\sim}{f}_1\right)^{-1} = h_{11} \tag{A3}$$

by well known properties of the inverse of a partitioned matrix.

b) Derivation of Equation (4.4)

By (4.3),

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = \left(I - K_t^{(P)}Z_t^{(A)}\right)T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + K_t^{(P)}\underset{\sim}{Y}_t^{(A)}. \tag{A4}$$

Hence,

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t = \left(I - K_t^{(P)}Z_t^{(A)}\right)\left(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t\right) + K_t^{(P)}\left(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t\right). \tag{A5}$$

The prediction errors $(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t)$ are independent of the residuals $(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t)$ and so,

$$P_t^{(A)} = E\left[\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)'\right] = Q_tP_{t|t-1}^{(A)}Q'_t + K_t^{(P)}\Sigma_t^{(A)}K_t^{(P)\prime} \tag{A6}$$

where we denote for convenience $Q_t = (I - K_t^{(P)}Z_t^{(A)})$.

By definition of the matrix $F_t^{(P)}$ (see below 4.3), equation (A6) can be written in the form

$$P_t^{(A)} = Q_t P_{t|t-1}^{(A)} - P_{t|t-1}^{(A)} Z_t^{(A)'} K_t^{(P)'} + K_t^{(P)} F_t^{(P)} K_t^{(P)'}$$

$$+ K_t^{(P)} \left(\Sigma_t^{(A)} - \Sigma_t^{(P)} \right) K_t^{(P)'} \tag{A7}$$

which implies the relationship (4.4) by straightforward algebra.

REFERENCES

ANDERSON, B.O.D., and MOORE, J.B. (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

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.

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

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, Vol. 6, (Eds.), P.R. Krishnaiah and C.R. Rao, Amsterdam: Elsevier Science, 187-211.

CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Analysis of Data in Time*. Ottawa: Statistics Canada (to appear).

COOLEY, T.F., and PRESCOTT, E.C. (1976). Estimation in the presence of stochastic parameter variation. *Economitrica*, 44, 167-184.

DIELMAN, T.E. (1983). Pooled cross-sectional and time series data: A survey of current statistical methodology. *The American Statistician*, 37, 111-122.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a State-Space model. *Journal of Econometrics*, 33, 388-397.

HARVEY, A.C. (1981). *Time Series Models*. Deddington, Oxford: Philip Allan.

HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

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.

JOHNSON, L.W. (1977). Stochastic parameter regressions: An annotated bibliography. *International Statistical Review*, 45, 257-272.

JOHNSON, L.W. (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. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 814-817.

MADDALA, G.S. (1977). *Econometrics*. Kogakusta: McGraw-Hill.

MEINHOLD, R.J., and SINGPURWALLA, N.D. (1983). Understanding the Kalman filter. *The American Statistician*, 37, 123-127.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 163-175.

PFEFFERMANN, D., and BARNARD, C. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* (to appear).

PFEFFERMANN, D., BURCK, L., and BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Analyisis of Data in Time*. Ottawa: Statistics Canada.

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.

ROSENBERG, B. (1973b). A survey of stochastic parameter regression. *Annals of Economic and Social Measurement*, 2, 381-397.

SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHWEPPE, F. (1965). Evaluation of likelihood functions for gaussian signals. *IEE Transactions on Information Theory*, 11, 61-70.

SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (Ed.) N.K. Nawboodivi, New York: Academic Press, 201-216.

SWAMY, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.

TILLER, R. (1989). A Kalman filter approach to labour force estimation using survey data. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association* (to appear).

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.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 239–253
Statistics Canada

A Method for the Analysis of Seasonal ARIMA Models

DAVID A. BINDER and J. PETER 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 and model validation can be performed. We apply these techniques to an unemployment series from the Labour Force Survey.

KEY WORDS: Kalman filter; Partial likelihood; Data smoothing.

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), Rao, Srinath and Quenneville (1989) and others considered the implications of certain stochastic models for the population means over time. Hausman and Watson (1985) incorporate a measurement error model into the standard seasonal adjustment process. Miazaki (1985) assumed that the survey error could be modelled with a pure moving average process. 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

[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.

smoothed values are also given, using the method described by Ansley and Kohn (1986). Bell and Hillmer (1987) used a similar model but their initial conditions do not extend easily to include regression terms or missing values (while preserving the marginal likelihood approach).

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 derive a smoothed estimate of the underlying process under the model assumptions. Recursive residuals are produced and validation techniques are used to evaluate the various models.

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:

$$\lambda(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \ldots - \lambda_P B^P,$$

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

$$\upsilon(B) = 1 - \upsilon_1 B - \upsilon_2 B^2 - \ldots - \upsilon_Q B^Q,$$

and

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

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

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

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; $b(B) = \upsilon(B^s)\beta(B)$, 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

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

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

and

$$a(B)u_t = b(B)\epsilon_t. \tag{2.5}$$

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 approximated by an autoregressive-moving average process, as suggested by Scott, Smith and Jones (1977).

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 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 $\lambda(B)$, $\alpha(B)$, $v(B)$ and of $\beta(B)$. The x_t's in 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 $m(0 \mid 0) = 0$ and covariance matrix $V(0 \mid 0)$, where

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

for large κ. The matrix $V_1(0 \mid 0)$ specifies the diffuse part of the prior. We explain in Section 3.2 how to obtain $V_1(0 \mid 0)$ and $V_0(0 \mid 0)$ for our model.

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

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

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 $m(t \mid t)$ and $V(t \mid t)$. Then the state vector z_t is augmented by the state $z_{t,r+1} = g_t' z_t$, and $m(t \mid t)$ and $V(t \mid t)$ 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, so that the last component of $m(T \mid T)$ 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 $V(t \mid t)$ gives the conditional variance. This procedure can be generalized to include any number of smoothed estimates and their conditional covariances. In applications, space limitations on the computer might preclude computing the smoothed values for a large number of time points.

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 $b = (b_1, \ldots, b_{q+sQ})$ with zeroes

to have dimension r, the ARIMA model may be written in the form given by (3.1), where $h_t' = (1, 0, \ldots, 0)$, $G_t' = (1, -b_1, \ldots, -b_{r-1})$ 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 $(r-1)$ by $(r-1)$ identity matrix and $0'$ is a row vector of zeroes.

In this formulation, the state vector $z_t = (z_{1t}, \ldots, z_{rt})'$ 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, 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 w_- is $N(0, \kappa I)$ and is independent of u_-. Since $(u_-', w_-')'$ is a non-singular linear combination of θ_-, the covariance matrix for θ 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 w_- is the null vector and we have $u_- = \theta_-$.

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 3.2, it is clear that e_t can be represented in state space form, with $h_t = (k_t, 0, \ldots, 0)'$, and $e_t = h_t'z_t$.

The regression component can be similarly represented by adding γ to the state vector and initially, assuming that γ has mean zero and covariance κI. Note that in the transition equation γ remains constant.

Since we can represent each of the components of y_t in expression (2.1) by a state space model, it is 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 $\lambda(B)$, $\alpha(B)$, $\upsilon(B)$ and $\beta(B)$. We transformed σ^2 to $\log(\sigma^2)$, in the numerical maximization procedure described below to avoid problems with negative parameter values. The model for the vector of observations $y = (y_1, y_2, \ldots, y_T)'$ given in Section 3 is equivalent to

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

where η is j-variate $N(\mathbf{0}, \kappa I)$, ζ is T-variate $N(\mathbf{0}, W)$, and M is some fixed $T \times j$ matrix. We note that η contains unknown constants including the regression coefficients; W is a function of the ARMA parameters; M is a function of the differencing structure.

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 this limit of the likelihood function is equivalent to the marginal likelihood function of $y - M\hat{\eta}$, where $\hat{\eta}$ is the maximum likelihood estimate of η when M and W are known. Tunnicliffe-Wilson (1989) has shown that the Jacobian of the transformation from the data y to $(\hat{\eta}, y - M\hat{\eta})$ does not depend on the model parameters of W whenever M is known. Ansley and Kohn (1985) have shown that M does not depend on the unknown parameters. By using the modified Kalman filter, the computations for the marginal likelihood function are more straightforward than the approach given by Tunnicliffe-Wilson.

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 $m(t \mid t')$ and the components of $V(t \mid t')$ 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 to choose the point 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 estimate the variance matrix for the estimated parameters, the inverse of the Fisher information matrix 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 as defined in (3.4) 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 $g'z_t$, it is sufficient to derive the variance $z_T - \hat{m}(T \mid T)$, where $\hat{m}(T \mid T)$ is the estimate of $m(T \mid T)$ at the estimated parameter values. This is because the state vector has been augmented to include $g'z_t$. Now,

$$z_T - \hat{m}(T \mid T) = [z_T - m(T \mid T)] + [m(T \mid T) - \hat{m}(T \mid T)]. \tag{4.2}$$

The first component of the right hand side of (4.2) has conditional variance $V(T \mid T) = V_0(T \mid T)$, assuming that $V_1(T \mid T) = 0$. 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

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

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

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

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).

4.3 Generalized Recursive Residuals

As Harvey and Durbin (1986) pointed out, useful quantities for performing model diagnostics are the generalized recursive residuals. In terms of our state space model, this is the difference between the observation and the one-step ahead prediction from the Kalman filter. These can be used for all time points t where $V_1(t + 1 \mid t) = 0$. Under the model, these residuals are approximately independent normal. They can be standardized to have an estimated variance of unity under the model. Diagnostics similar to those used in classical regression models can then be performed.

5. ANALYSIS OF LABOUR FORCE DATA

5.1 Parameter Estimation

To demonstrate this procedure, we take data from the Canadian Labour Force Survey (LFS). The LFS is a monthly rotating panel survey with each panel containing one-sixth of the selected households. A panel will remain in the sample for six consecutive months while the primary sampling units will rotate out after approximately two years. The sample selection follows a stratified multi-stage design.

The data were the monthly number of unemployed as published from January 1977 to December 1986 for the province of Nova Scotia and for the subprovincial region within Nova Scotia corresponding to Cape Breton Island. This province was selected because the sampling errors are moderate compared to the larger provinces. Cape Breton Island was selected because its smaller sample size provides estimates with a larger relative variance. Graph 1a displays the logarithm of the Nova Scotia series and Graph 1b shows the similarly transformed Cape Breton Island series. We used the logarithms as our inputs.

Lee (1990) estimated the autocorrelations for the Nova Scotia survey error up to a lag of eleven. We derived the coefficients of the ARMA (m,n) survey error process given in (2.7) by matching these autocorrelations. A good fit was found using an ARMA (3,6) model. The resulting coefficients were:

$$\begin{aligned} \phi_1 &= 0.2575 & \psi_1 &= -0.1847 \\ \phi_2 &= -0.3580 & \psi_2 &= -0.5873 \\ \phi_3 &= -0.6041 & \psi_3 &= 0.3496 \\ & & \psi_4 &= 0.0647 \\ \tau^2 &= 0.7246 & \psi_5 &= 0.0982 \\ & & \psi_6 &= 0.0347. \end{aligned}$$

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.

A series of models were fitted to the Nova Scotia data with an assumption of no sampling error. The same models were then refitted, incorporating the model for the survey error process. In this case we could also compute smoothed values for the survey estimates and compare their standard errors with the standard errors of the original series.

The preliminary model selected for the Nova Scotia data, ignoring the sampling error, was a seasonal ARIMA $(1,1,0)(0,1,1)_{12}$. However the moving average term for the seasonal component was estimated to be one, so a deterministic regression term was used 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 appropriate for this model because the first differences of the data are fitted.

Further analysis of this reduced model showed that the moving average seasonal component was not required in the model. The final model selected for the Nova Scotia data was an ARIMA (1,1,0) with a deterministic regression component. This same model was then used for the Nova Scotia data with the survey error process incorporated. The same structural model was used for the Cape Breton Island series.

Table 1 displays the parameter estimates. The estimates that do not incorporate the survey error component are in the **Without Sampling Errors** columns. First, examining the models for Cape Breton Island shows that the regression estimates are similar, as would be expected. Note that the autoregressive estimates (AR) are also similar and that the **With Sample Error** model has reduced the estimated model variance substantially. The column headed ***T*-value** displays the estimated parameter divided by its standard error. Note that the t-values for the autoregressive parameter are substantially different (-0.68 vs -2.85). This would lead to

Table 1

Parameter Estimates - Unemployment Series 1977-1986

Parameter	Nova Scotia				Cape Breton Island			
	Without Sampling Error		With Sampling Error		Without Sampling Error		With Sampling Error	
	Estimate	*T*-value	Estimate	*T*-value	Estimate	*T*-value	Estimate	*T*-value
Alpha	−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

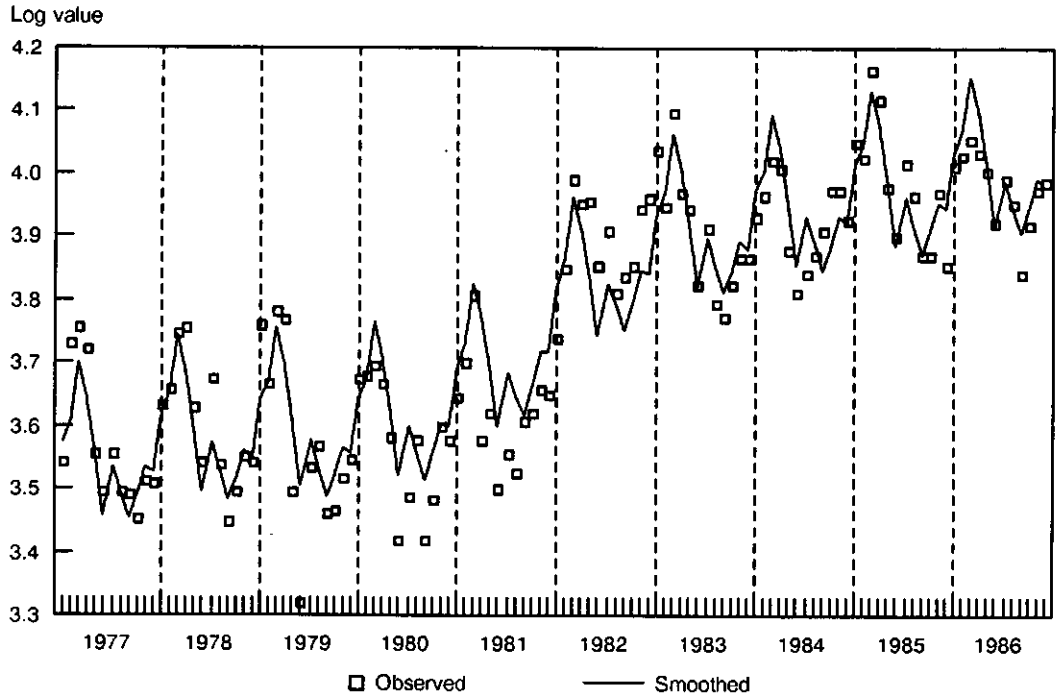
accepting a model for the Cape Breton Island data with only a deterministic regression term when the survey error process is incorporated into the model. However, if the survey error is ignored in the analysis, too much significance would be attached to the autoregressive parameter.

The results for the Nova Scotia models are also displayed on Table 1. Note that the reduction in the estimate of the model variance by incorporating the sampling error structure is much greater for the Nova Scotia series than was achieved for the Cape Breton data. An important result in the Nova Scotia models is the difference in the estimates for the autoregressive component. Both models show that the AR component is highly significant in each model. The **Without Sample Error** model gives an estimate of $\alpha = -0.296$; whereas the **With Sample Error** model gives an estimate of $\alpha = 0.862$. Clearly, the interpretations that would be associated with these two estimates are entirely different.

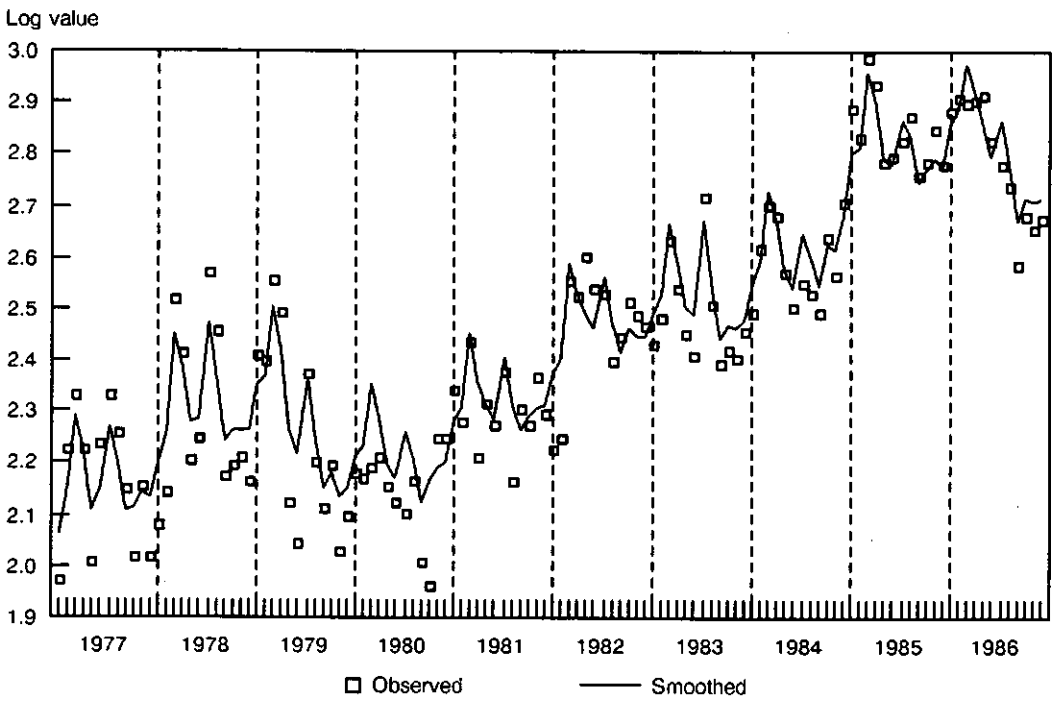
The smoothed estimates for the model incorporating sampling error are shown superimposed on the original data series in Graph 1a. Graph 1b shows the smoothed estimates for Cape Breton Island superimposed on the original series. The most notable item in these plots is the impact of the recession of 1981 on the smoothed estimates. Prior to the recession, the model tends to overestimate unemployment and after 1981 the model tends to underestimate the number of unemployed.

5.2 Model Validation

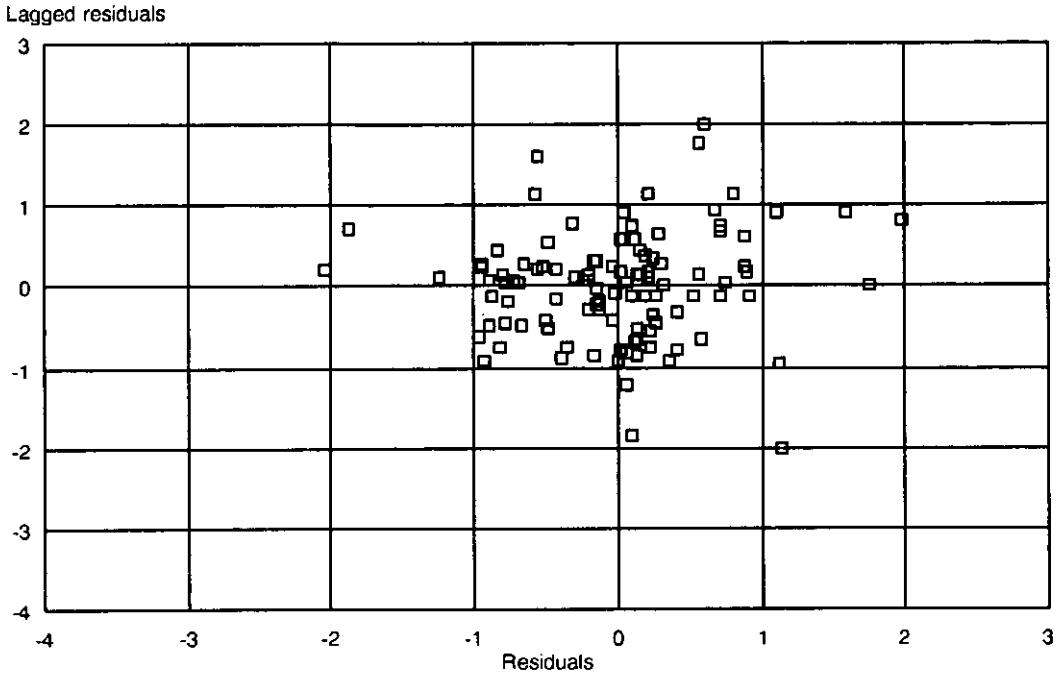
The plots of the generalized recursive residuals (described in Section 4.3) against the lagged generalized recursive residuals were produced for all the models. Graphs 2a and 2b show these plots for the two models for Nova Scotia. Note that Graph 2a shows less dispersion around the origin than Graph 2b, indicating a better fit when survey error is incorporated in the model.



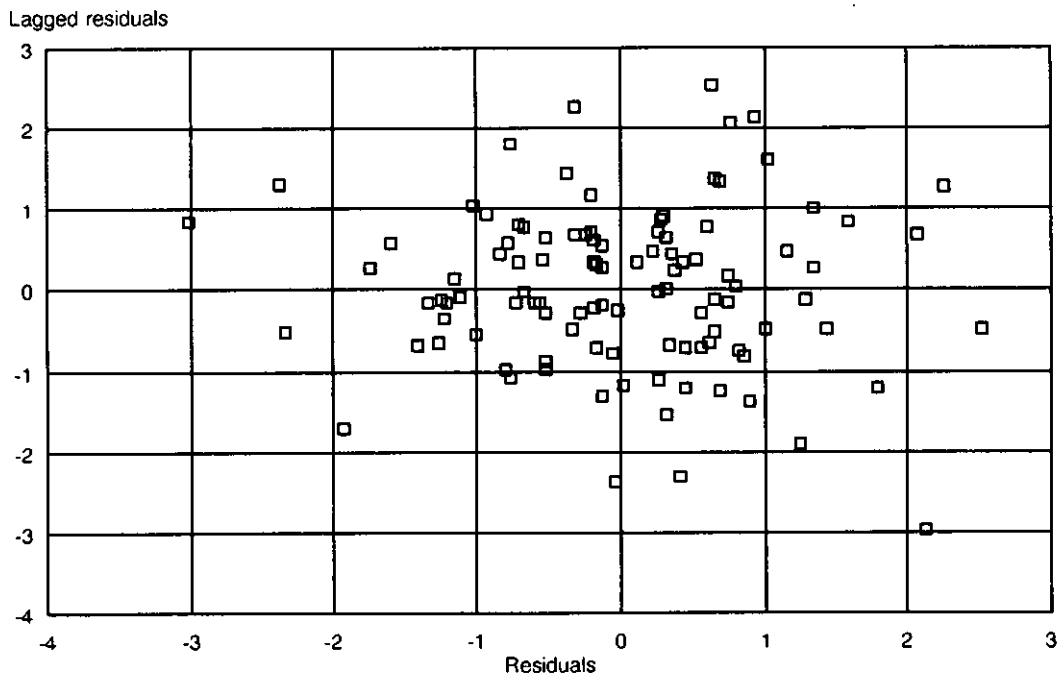
Graph 1a Nova Scotia Observed and Smoothed Values (Log Transform)



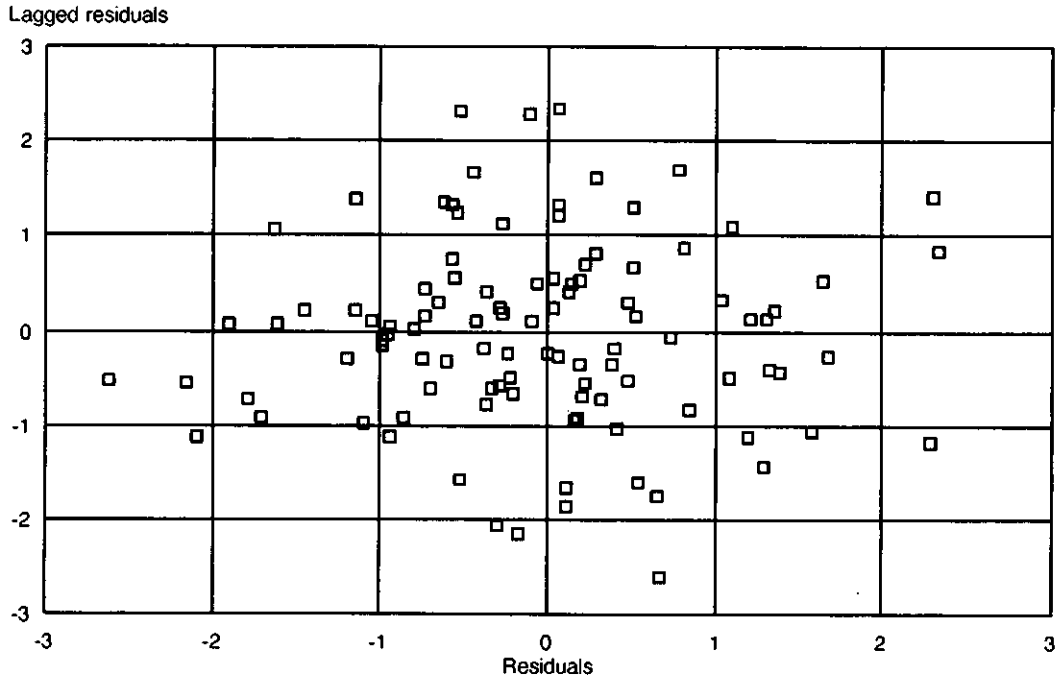
Graph 1b Cape Breton Island Observed and Smoothed Values (Log Transform)



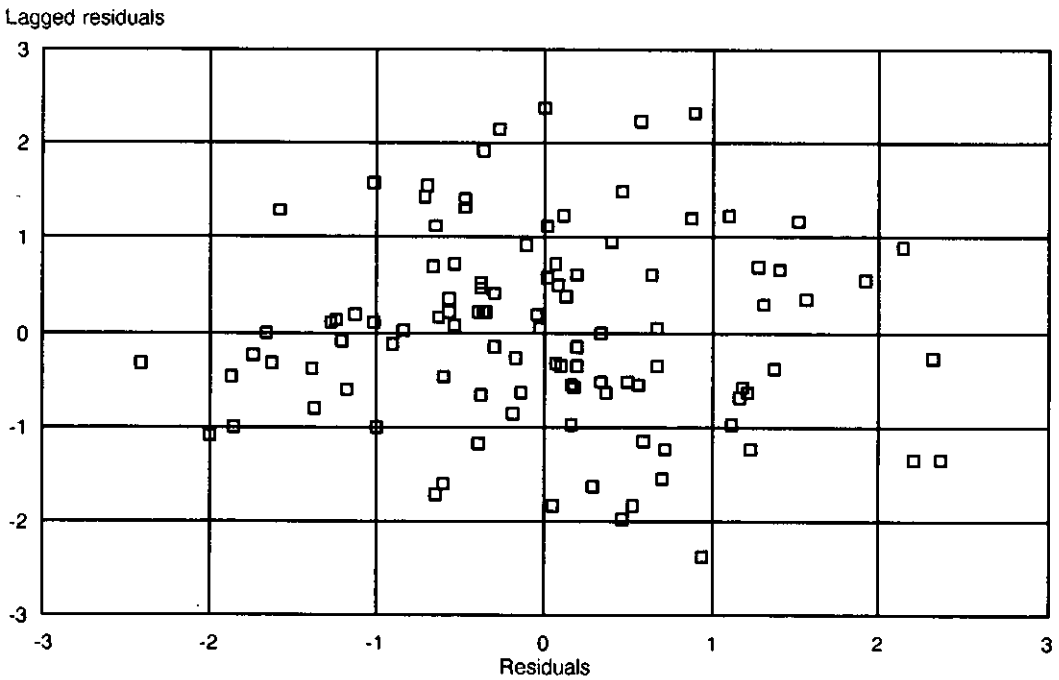
Graph 2a Nova Scotia One Step Ahead Prediction Errors – Survey Error Included



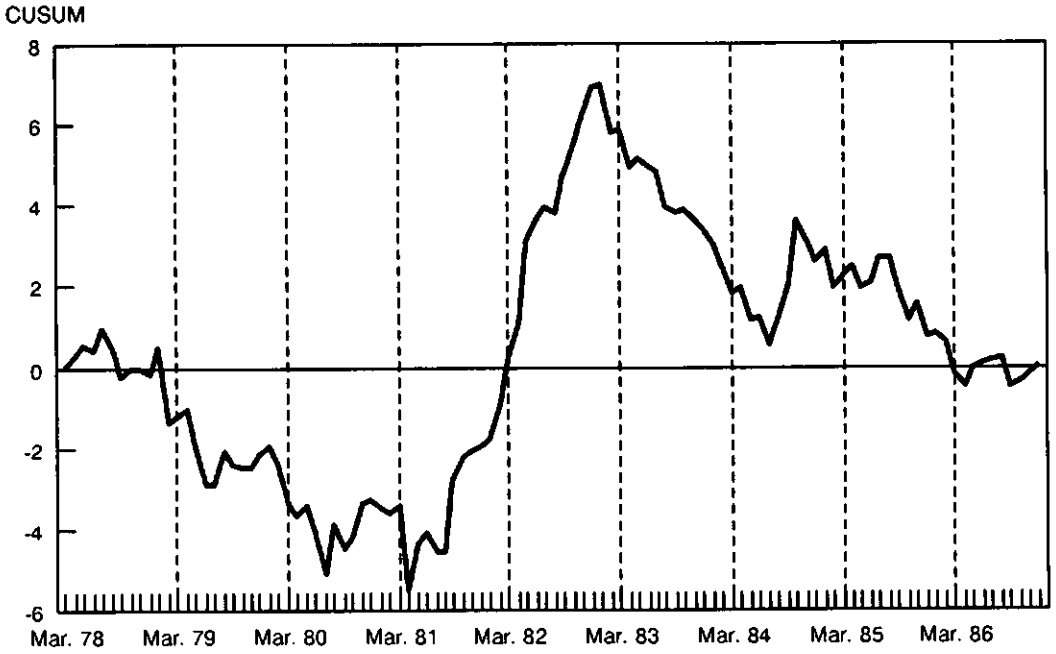
Graph 2b Nova Scotia One Step Ahead Prediction Errors – Survey Error Ignored



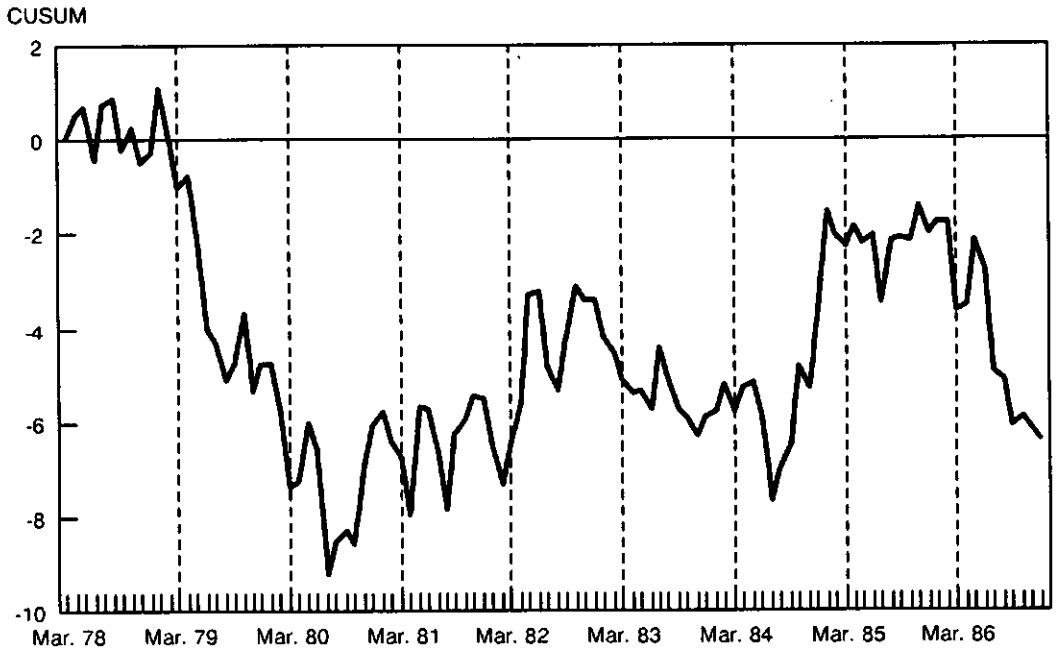
Graph 3a Cape Breton Island One Step Ahead Prediction Errors – Survey Error Included



Graph 3b Cape Breton Island One Step Ahead Prediction Errors – Survey Error Ignored



Graph 4a Nova Scotia CUSUM of One Step Ahead Prediction Errors



Graph 4b Cape Breton Island CUSUM of One Step Ahead Prediction Errors

The same plots for Cape Breton Island are shown in Graph 3a and 3b. There is a striking similarity in the resulting residual plots for the two models from Cape Breton. However, none of the four plots give any compelling reason to doubt the underlying normal assumption of any of the models.

To test that the models did not undergo a structural change, the recursive residuals can be cumulatively summed to create a CUSUM chart. Whereas using the tests described in Brown, Durbin and Evans (1975) produced no significant results, the chart does suggest some structural change may be occurring. The CUSUM for Nova Scotia, as displayed in Graph 4a, shows quite clearly that prior to the recession the residuals are generally negative, implying that the model predictors are too large. During the 1981 recession the model produces mainly positive residuals. This implies that the model predictors are too small. The CUSUM for the Cape Breton Island models is shown in Graph 4b. Here we can see that the model that includes the survey error undergoes an earlier structural change.

We see, therefore, that model improvements can be made. By incorporating an extra regression variable corresponding to the structural changes noted in the CUSUM chart, further analysis can be performed within the same general framework. The form of such a variable is currently being investigated.

5.3 Summary

These examples demonstrate the importance of accounting for survey errors in certain time series analyses. Using the modified Kalman filter, we have developed a flexible method for parameter estimation, data smoothing and model validation for a wide variety of commonly used models.

ACKNOWLEDGEMENTS

The authors are grateful to Bill Steele of Social Survey Methods Division for his work in programming the algorithm described in Section 4. We are also grateful to the Labour Force Survey for supplying the data series. The referee and the associate editor both supplied many useful comments that improved the paper.

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.

BELL, W.R., and HILLMER, S.C. (1987). Time Series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.

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. *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

BROWN, R.L., DURBIN, J., and EVANS, J.M. (1975). Techniques for testing the consistency of regression relationships over time. *Journal of the Royal Statistical Society*, Series B, 37, 149-163.

HARVEY, A.C., and DURBIN, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society*, Series A, 149, 187-222.

HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

HAUSMAN, J.A., and WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

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. (1990). Estimation of panel correlations for the Canadian Labour Force Survey. *Survey Methodology*, 16, 283-292.

MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. *Ph. D. Thesis*, Iowa State University, Ames, Iowa.

RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kaltin and M.P. Singh), New York: Wiley, 457-479.

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.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 255–269
Statistics Canada

Spatial-Temporal Modelling of Spatially Aggregate Birth Data

DAVID R. BRILLINGER[1]

ABSTRACT

Births by census division are studied via graphs and maps for the province of Saskatchewan for the years 1986-87. The goal of the work is to see how births are related to time and geography by obtaining contour maps that display the birth phenomenon in a smooth fashion. A principal difficulty arising is that the data are aggregate. A secondary goal is to examine the extent to which the Poisson-lognormal can replace for data that are counts, the normal regression model for continuous variates. To this end a hierarchy of models for count-valued random variates are fit to the birth data by maximum likelihood. These models include: the simple Poisson, the Poisson with year and weekday effects and the Poisson-lognormal with year and weekday effects. The use of the Poisson-lognormal is motivated by the idea that important covariates are unavailable to include in the fitting. As the discussion indicates, the work is preliminary.

KEY WORDS: Aggregate data; Borrowing strength; Contouring; Extra-Poisson variation; Locally-weighted analysis; Maps; Periodogram; Poisson distribution; Poisson-lognormal distribution; Random effects; Spatial data; Time series; Unmeasured covariates.

1. INTRODUCTION

The concern of this work is spatial-temporal data, that is quantities recorded as functions of space and time. The analysis of such data should be "easy" because of the graphing possibilities, *e.g.* rate versus time or effect versus geography, in the manner of residual plots so often employed in regression analysis; however in the present case the aggregation of basic elements leads to substantial difficulties.

The specific data studied consists of daily births for the calendar years 1986 and 1987 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 fairly regular. (The latter was important at the early stages of the work because computer based maps were then unavailable). Women aged 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. They are characterized by being aggregate, by being non Gaussian and by being non stationary in space and time.

It is wished to understand the relationship of births to time and geography, specifically to allow temporal and spatial patterns of fertility and possible surprises to show themselves. There are two central aspects to the study; a locally-weighted analysis of aggregate data is developed and random effects models are set down and fit to handle extra-Poisson variation. The latter part may be viewed as an inquiry into the flexibility of the Poisson-lognormal to handle unmeasured covariates and errors. The locally weighted analysis proceeds by developing weights, $w_i(x,y)$, that are meant to reflect the influence of the i-th census division (an aggregate) on the point location with coordinates (x,y). Given census division data, these

[1] David R. Brillinger, Statistics Department, University of California, Berkeley, CA, 94720, U.S.A.

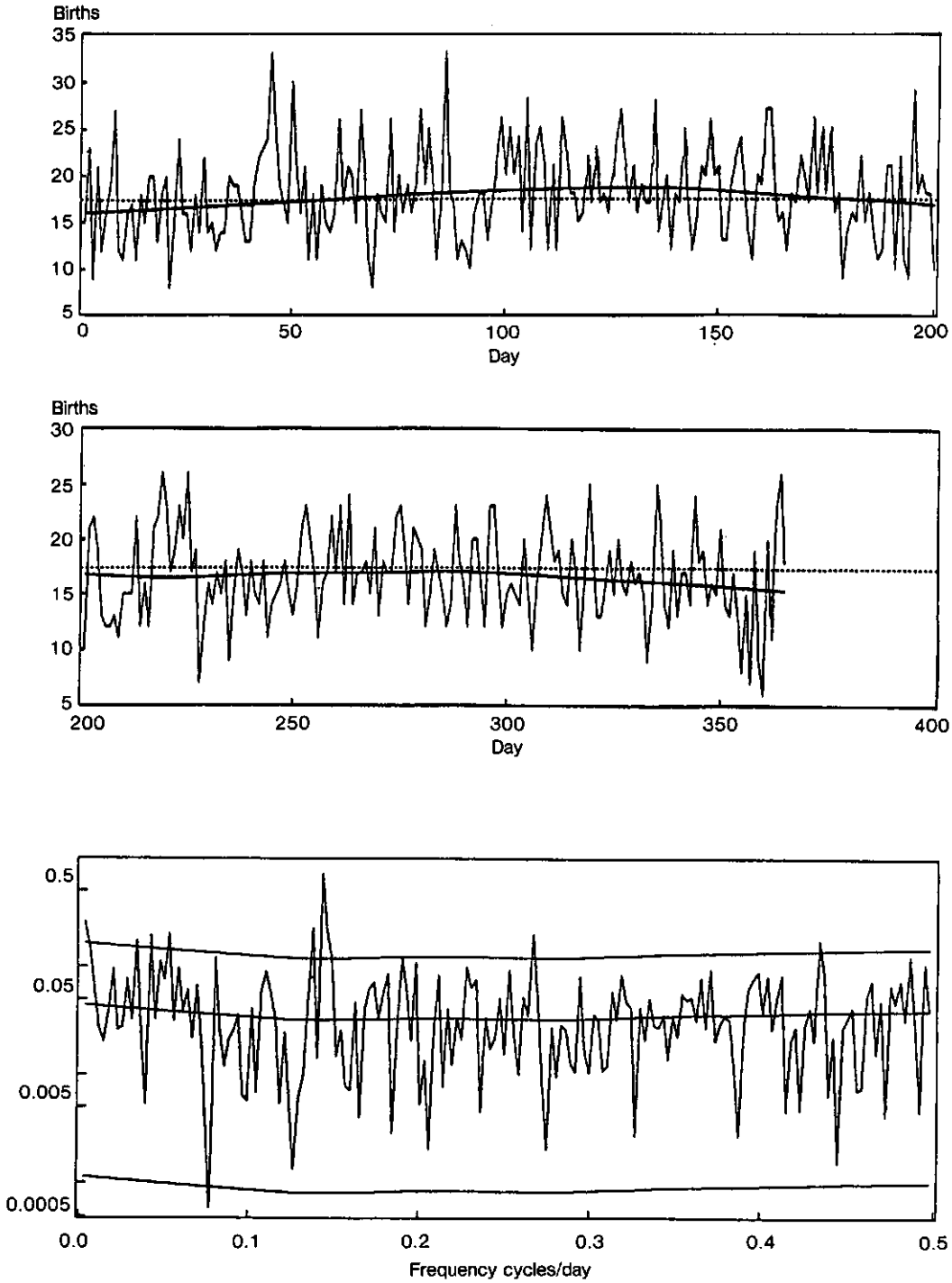


Figure 1. Top: Time series of annual births to women aged 25-29 in 1986 for the Province of Saskatchewan. Bottom: Periodogram of the square roots of the count graphed above. The solid lines provide approximate 95% marginal confidence limits. The peak corresponds to a period of 7 days.

weights are then applied to individual terms of the log-likelihood or corresponding estimation equations and parameter estimates evaluated.

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 expressions employed for the weights, in this present work, are naive and bound to change form with further study, but the character of the analysis may be anticipated to remain of some interest.

The companion paper Brillinger (1990) considers some aspects of the spatial case alone.

2. BIRTHS AS A TIME SERIES

The top graph of Figure 1 provides the total number of births in Saskatchewan for each day of 1986. The dashed line is the 1986 mean level. The solid line is the result of heavily smoothing the series and is meant to highlight any trend. This graph does not, with casual inspection, provide striking evidence of any special phenomenon. However when the periodogram of the square root of the counts is computed, see bottom graph of Figure 1, something of interest appears. (The square root is employed to make the series more nearly symmetrical and normal). The upper and lower solid lines on the graph provide approximate 95% marginal confidence limits about a heavily smoothed version. A peak is apparent at a frequency of .143 cycles/day corresponding to a period of 7 days. This periodic phenomenon is well known in the analysis of birth data, see *e.g.* Cohen (1983) and Miyaoka (1989) and references therein. It is usually ascribed to doctors intervening in the natural process of labour and inducing births particularly on weekdays.

3. BIRTHS AS A SPATIAL PROCESS

Figure 2 provides, for each census division, and for women aged 25 to 29 the annual rate of births for the years 1986 and 1987 combined. One sees the highest rate of .208 births per woman per year to occur in the northern half of the province while the two lowest rates appear in the census divisions containing Regina and Saskatoon.

Figure 3 provides the numerical difference between the annual rate for 1987 and that for 1986 for each of the 18 census divisions. (Note that the 1986 census population has been taken as the divisor in each case). The differences are scattered around 0. It is to be noted that these rates are, however, based on fairly widely varying population sizes.

In the previous section the presence of a phenomenon of period 7 days was noted. Figure 4 presents the difference between the average weekday rate and the average weekend rate, (weekdays meaning Monday through Friday) for each census division. In all but one census division, the weekday rate is higher. This is consistent with various other studies and, as suggested in Section 2, is very possibly due to doctors inducing labor on weekdays (to avoid births on weekends).

The various rates presented in Figures 2, 3, 4 are average values for individual census divisions.

4. PROBLEMS ARISING

Maps of most quantities of direct interest that assign average values to the wholes of counties thereby lie, lie, lie.

With these graphic words Tukey (1979) deplores the use of maps such as those of Figures 2, 3, 4 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. A principal concern of this work is to develop contour maps displaying smooth variation. It is hoped that such maps will prove useful in the discovery of general stochastic descriptions of the phenomenon and will allow insightful exploratory analyses.

A second concern of this work is with the statistical distribution of the counts themselves. A natural special stochastic model to employ is the Poisson. Yet in past studies the birth process has been found to relate to many socio-economic quantities, *e.g.* diet, lifestyle, weather, environment, weekday, holidays, age structure. Further the population of the various census divisions has varied around the Census Day values throughout 1986-87 and lastly the women's ages are scattered from 25 to 29. In summary it seems necessary to employ a more flexible model than the Poisson, specifically 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, in the manner described by Mallows and Tukey (1982). (The term "borrowing strength" is employed, rather than for example "empirical Bayes" as some might prefer, because it has been in use for a substantial time period and because of its broader implications). Dean *et al.* (1989) is another recent reference concerned with handling extra-variation.

5. 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 log-likelihood

$$\sum_i W_i(x,y) \log p(Y_i \mid \Theta) \tag{1}$$

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 at (x,y) results from minimizing

$$\sum_i W_i(x,y)\, [Y_i - \mu]^2$$

and is given by

$$\hat{\mu}(x,y) = \sum_i W_i(x,y)\, Y_i \Big/ \sum_i W_i(x,y),$$

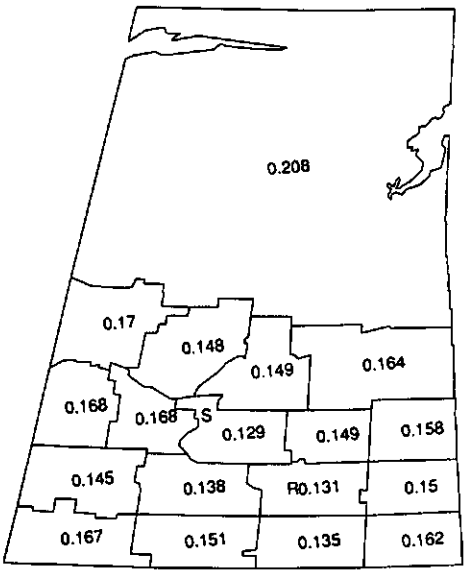


Figure 2. The average annual birth rate for women aged 25 to 29 for the years 1986 and 1987, plotted above census divisions. "R" and "S" indicate the locations of Regina and Saskatoon respectively.

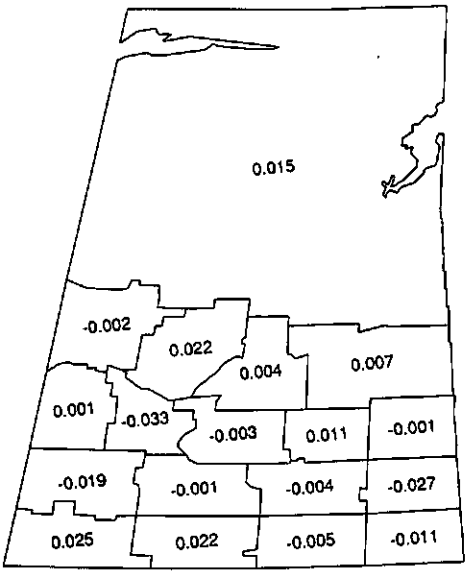


Figure 3. The 1987 rate minus the 1986 rate for the same data as Figure 2.

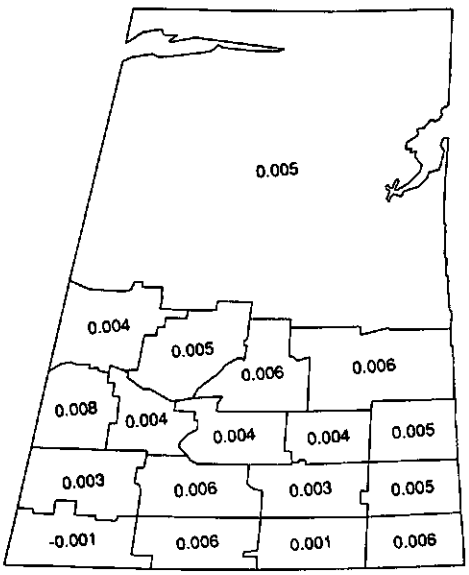


Figure 4. The average weekday rate minus the average weekend rate for the same data as Figure 2.

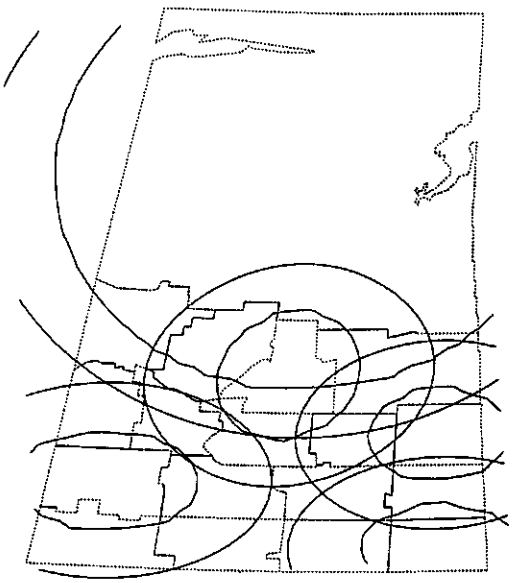


Figure 5. The weights, $W_i(x,y)$ applied in equations (1) or (2), computed via expression (4), for four of the census divisions. The weights are not shown for all the divisions in the interests of clarity. The contours at levels .50 and .99 are shown.

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. Specifically consider the loss function

$$L(Y \mid \Theta) = -\log p(Y \mid \Theta).$$

Suppose an estimate is desired at $\mathbf{r} = (x,y)$. The Bayes' risk may be written

$$E\{L(Y \mid \Theta_{\mathbf{r}})\} = E\{E\{L(Y \mid \Theta_{\mathbf{r}}) \mid \mathbf{r}\}\}.$$

Bayes' rule seeks

$$\min_{\Theta} E\{L(Y \mid \Theta) \mid \mathbf{r}\}.$$

With data Y_i, $\mathbf{r}_i$, and $W_i(\mathbf{r})$ a kernel centred at $\mathbf{r}_i$, one approximates the conditional expected value here by

$$E\{\log p(Y \mid \Theta) \mid \mathbf{r}\} \approx \sum_i W_i(\mathbf{r}) \log p(Y_i \mid \Theta)$$

and so is led to expression (1).

Tibshirani and Hastie (1987) develop an equi-weighted local likelihood estimation procedure. Cleveland and Devlin (1988) develop the least squares approach in real detail. 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 non-additivity, variants for resistance and influence, simple additivity of the observation component, and no matrix inversion (as, for example, kriging requires).

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

$$w_i(x,y) = \frac{1}{\mid R_i \mid} \quad \text{for } (x,y) \text{ in } R_i$$

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 $W(\cdot)$ is a kernel appropriate for the nonaggregate case as for example studied in Cleveland and Devlin (1988). The formula (3) may be motivated by consideration of the Poisson point process case, see Appendix II. Estimates will be determined via the criteria (1) or (2) with W_i replaced by w_i.

The specific weights employed at $\mathbf{r} = (x,y)$ in this preliminary work are

$$w_i(\mathbf{r}) = \exp\{-(1-\rho)^2 \|\mathbf{r} - \mathbf{r}_i\|^2/2\tau^2\} \tag{4}$$

outside the ellipse $(\mathbf{r}_0 - \bar{\mathbf{r}}_i)\mathbf{S}_i^{-1}(\mathbf{r}_0 - \mathbf{r}_i)' = d_0^2 = 5.991$ and equal 1 inside. Here $\|\mathbf{r}\|^2 = x^2 + y^2$, $\rho = d_0/\sqrt{(\mathbf{r} - \bar{\mathbf{r}}_i)\mathbf{S}_i^{-1}(\mathbf{r} - \bar{\mathbf{r}}_i)'}$ and $\tau = .025$, where $\bar{\mathbf{r}}_i = E\,U_i$ and $\mathbf{S}_i = \operatorname{var} U_i$ with U_i a variate uniformly distributed within R_i. This choice of ρ makes the weight function continuous. 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). One could have employed other shapes than ellipses, *e.g.* rectangles, but this is preliminary work and it is anticipated that later work will employ weights of the form (3).

Figure 5 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. The jaggedness in some of the contours results from the discreteness of the 40 × 40 grid employed in the computations.

Other weight functions constructed with somewhat 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 in (1) or (2) 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 other counties. The concern of this present paper however is the influence of a "county" on a point location. It is to be remarked that perhaps the weight, providing the influence, should depend on some covariates, *e.g.* county population.

6. A POISSON FIT

Throughout the analysis, the female population aged 25-29 and births to its members will be considered. Let $i = 1, \ldots, 18$ 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 two years 1986-87.

Suppose that the probability distribution $p(\cdot)$ of Section 5 is that B_i is Poisson with mean $2N_i\mu$. (The presence of the multiplier 2 is so the parameter μ is an annual 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 annual birth rate at location (x,y) is given by

$$\hat{\mu}(x,y) = \sum_i w_i(x,y)\,B_i \Big/ 2\sum_i w_i(x,y)\,N_i. \tag{5}$$

These values are computed for (x,y) on a 40 by 40 grid and the corresponding contour plot is given in Figure 6. The contours are seen to vary smoothly. This (smoothed) rate varies from .14 to .20, with the higher values in the upper half and the lower centred around the Province's most urban part.

As indicated previously, the data under study has important temporal characteristics. Models need to take this into account. In particular the weekly periodicity needs to be handled as well as possible trends in population sizes. The following model seems worth considering. Let j be an indicator variable with $j = 1$ if the count is for a weekday and $j = 2$ if the count is for

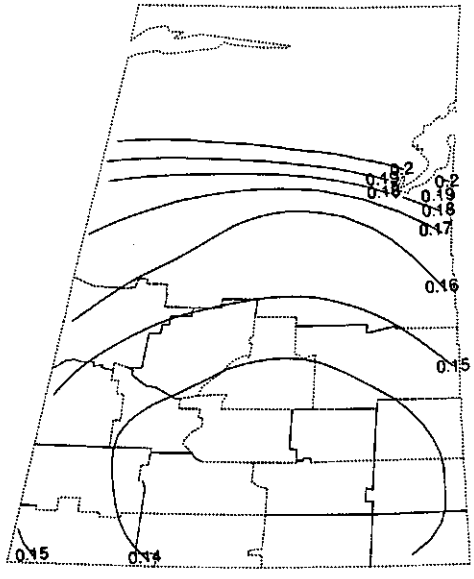


Figure 6. Expression (5) graphed for the weights of (4) with B_i the count of births in census division i during 1986-87 and N_i the corresponding population count of women aged 25-29.

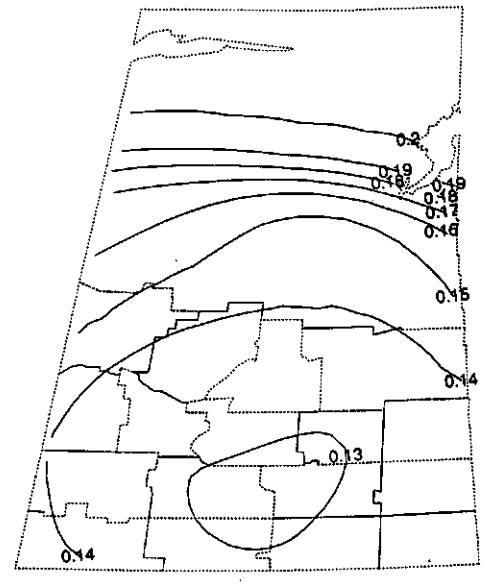


Figure 7. The estimated birth rate $\exp\{\hat{\alpha}\}$ obtained by locally weighted fitting assuming that the number of births, B, given the population at risk, N, is Poisson with mean $N\exp\{\alpha \pm \beta \pm \gamma\}$ with the first $\pm$ sign plus for weekdays and minus for weekends and the second $\pm$ plus for 1986 and minus for 1987.

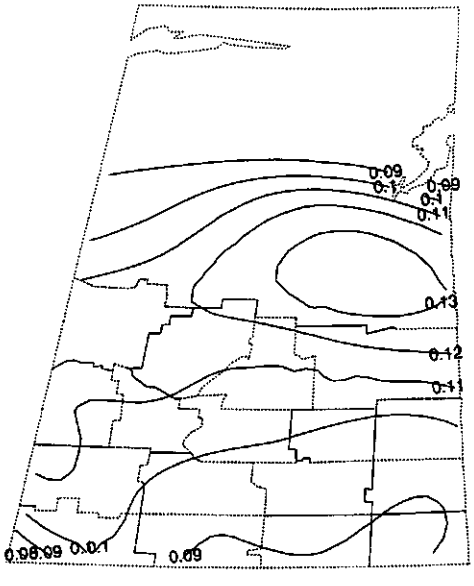


Figure 8. Plot of the estimated weekday effect $\hat{\beta}(x,y)$ obtained as per Figure 7.

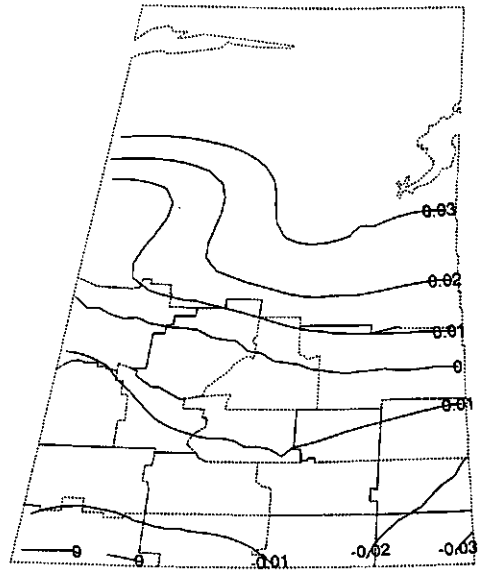


Figure 9. The estimated year effect $\hat{\gamma}(x,y)$ as per Figure 7.

a weekend. Let k be a second indicator variable with $k = 1$ for 1986 and $k = 2$ for 1987. Let B_{ijk} denote the corresponding number of births in census division i. Suppose that B_{ijk} given N_i is Poisson with mean $N_i \exp\{\alpha + \beta_j + \gamma_k\}$. β_j is the weekday effect, γ_k the year effect and it will be assumed that $\beta_1 + \beta_2, \gamma_1 + \gamma_2 = 0$ to make the model identifiable. If there is no weekday effect, then $\beta_1, \beta_2 = 0$. If there is no year effect, then $\gamma_1, \gamma_2 = 0$. Now, via locally-weighted analysis presented in Section 5, one can obtain estimates of α, β and γ as functions of location (x,y). (For simple balance in the computations, only the first $364 = 7 \times 52$ days of each year have been employed).

Figure 7 provides the estimate $\exp\{\hat{\alpha}(x,y)\}$ obtained of the annual birth rate. It is interesting to note that, relative to the constant rate Poisson model, the contours have expanded out somewhat from the urban areas. Figure 8 provides the estimated weekday effect, $\hat{\beta}_1(x,y)$, obtained. In its case there is bulge to the east. These values are quite a different representation from that of the naive differences of Figure 4. In particular, now there is a reflection of the differing population sizes. The order of magnitude of the $\hat{\beta}$'s is .08 to .13 while $\hat{\alpha}$ is order -2.1 to -1.6. Figure 9 provides the estimated year effect, $\hat{\gamma}_1(x,y)$. Its values vary from $-.03$ to .03. Numerically, the weekday-weekend effect is the larger.

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. POISSON-LOGNORMAL FITS

With a multi-dimensional explanatory variable $\mathbf{x}$ in hand, a Poisson model that has B of mean $N\exp\{\mathbf{x}\Theta\}$ 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, the variate B is Poisson with mean $N\mu\exp\{\epsilon\}$ and that ϵ is normal with mean 0 and variance σ^2. In the case of this model B is said to have a Poisson-lognormal distribution. Some information on this distribution may be found in Shaban (1988). Sometimes ϵ enters directly from the problem context, see Brillinger and Preisler (1983) for one example, but in the present case it is simply assumed present.

A critical difficulty, that arises in working with a Poisson-lognormal model, is that closed expressions do not exist for the probability function. Yet the model is clearly flexible for introducing effects and handling unavailable variables. Following the work of Bock and Lieberman (1970), Pierce and Sands (1975) and Hinde (1982), one can proceed via numerical quadrature. 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 $N\mu$. To proceed with a data analysis the integral is approximated by a finite number of terms involving nodes, z_l, and weights, w_l,

$$p(Y) \approx \frac{1}{Y!}\sum_{l=1}^{L} (ve^{\sigma z_l})^{Y}\exp\{-ve^{\sigma z_l}\}w_l.$$

Listings of nodes and weights may be found in Abramowitz and Stegun (1964) for example.

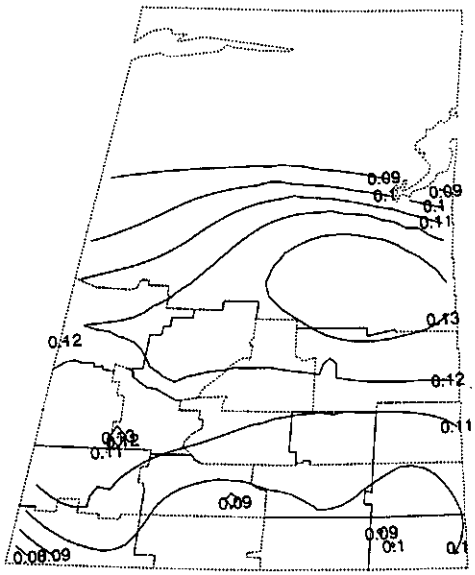


Figure 10. A plot comparable to Figure 7, except that now a normal error term is added to the linear predictor.

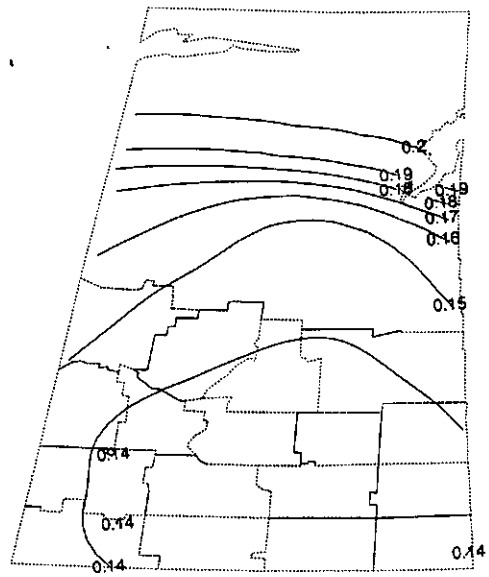


Figure 11. A plot comparable to Figure 8, except now (as in Figure 10) a normal error term has been added to the linear predictor.

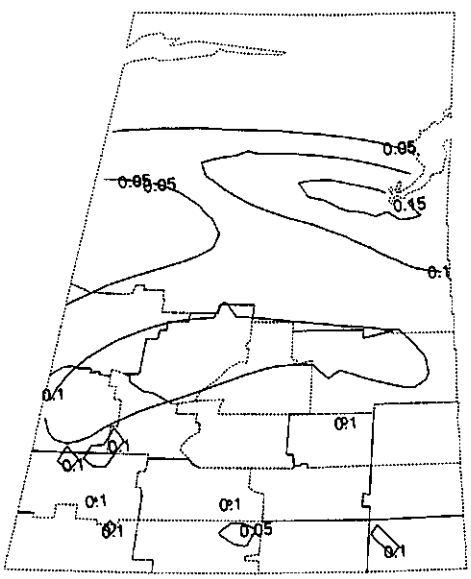


Figure 12. A plot comparable to Figure 9, except now (as in Figure 10) a normal error term has been added to the linear predictor.

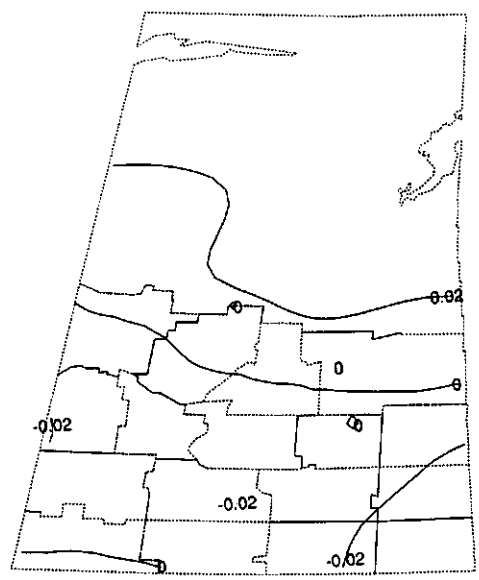


Figure 13. The estimated standard error, $\hat{\sigma}(x,y)$, of the normal term added to the linear predictor.

Figures 10, 11, 12, 13 provide the results of fitting the Poisson-lognormal model including weekday and year effects and employing $L = 5$ nodes. The model assumes B_{ijk} given N_i and Z is Poisson with mean

$$N_i \exp\{\alpha + \beta_j + \gamma_k + \sigma Z\}$$

Z denoting a standard normal deviate and further assumes the separate Z's independent. Here i indexes census division, j weekday or not and k year. Figure 10, a contour plot of $\exp\{\hat{\alpha}(x,y)\}$, again shows a dip around the urban region as in Figure 7. The irregularity in the figure suggests that in one case perhaps the estimation procedure converged to a local extremum. Figures 11 and 12 similarly provide $\hat{\beta}(x,y)$ and $\hat{\gamma}(x,y)$. There are again suggestions of local extrema. Figure 13, a contour plot of $\hat{\sigma}(x,y)$, is not easily described. It suggests that the estimate, $\hat{\sigma}$, is fairly variable. The estimate is seen to be of order of magnitude .1 and so comparable to the weekday effect of Section 6.

All the work on estimation with the Poisson-lognormal, that we know about, involves some form of approximation. For example Clayton and Kaldor (1987) approximate the conditional Poisson log-likelihood by a quadratic and Aitchison and Ho (1989) also employ numerical integration, albeit after a transformation of the parameters. A new type of approximation has recently been proposed in Crouch and Spiegelman (1990). Its effectiveness for the Poisson-lognormal remains to be studied.

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 omitted 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 manageable on the Berkeley network of Sun 3/50's.

Much future work remains including: tools for assessing fit, uncertainty computation and display, weight function choice (particularly choice of τ in (4)), analyses for other age groups and provinces, and appropriate asymptotics. Further understanding needs to be gained as to why with nearby initial values the optimizing routine sometimes converged to somewhat distant estimates. An advantage of the present circumstance is that there exists immense amounts of other data to be made use of as work progresses. Examination of Figures 6 on shows an important limitation of the technique – it is providing too much fine detail in the northern half of the province.

Other recent papers devoted to the analysis of vital statistics rates are: Cressie and Read (1989), Clayton and Kaldor (1987), Tsutakawa (1988) and Manton *et al.* (1989). These papers are however not directed at the problem of obtaining a smooth surface, which is the concern of this work.

It is amusing to note that the presence of the weekly period in the phenomenon allowed the author to deduce early on in the work that a confusion had arisen over which data set was to be supplied. When the days of fewest births were determined for the initial data set supplied, the days were found to be (apparently) Friday and Saturday. This was because the year 1987 had been supplied, and not the desired 1986.

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. Luckily the boundaries have not changed much, but this circumstance provides yet more reason for wanting a procedure that can handle extra-variation.

9. ADDENDUM

In the paper a case has been made for the inclusion of an error term, ϵ, to reflect pertinent covariates that were unavailable for the analysis. This led to the employment of the Poisson-lognormal distribution. In Tukey (1990) an index of urbanicity of a census division is constructed. It is based on the populations of the three largest places in the division. The values, x_i, of the index are given in Figure 14 and are seen to be lowest in the census divisions containing Regina and Saskatoon.

The table below gives the results of employing Glim to fit the successive Poisson models for B_{ijk} given N_i: (i) $N_i \exp\{\alpha + \beta_j + \gamma_k\}$, (ii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta x_i\}$, and (iii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta_1 x_i + \delta_2 x_i^2\}$.

Variables	Deviance	d.f.	p-value
weekday, year	227.3	69	
+ urbanicity	86.69	68	
+ urbanicity**2	83.13	67	.088

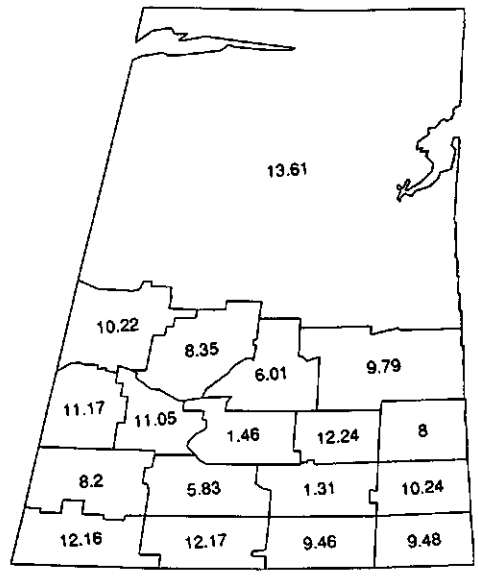


Figure 14. The values of the Tukey index of urbanicity.

By bringing in this urbanicity variable, x_i, now a Poisson model is satisfactory for the circumstance.

Finally the Referee made some comments that spell out quite specifically the assumptions and limitations of this present study. The work is continuing and the intention is to address these comments. Rather than paraphrasing, it seems more sensible to provide the referee's own words.

"The choice of weights is *ad hoc* and requires more thought. If one had two divisions, both of the same area but with vastly different populations N_i, should the weighting be the same? It depends on whether area or population density is thought to be more important. Use of the latter may remove the spurious fine detail in the northern half of the province."

"There are traps with N_i's, which the author appears to be aware of, but I think the reader needs extra warning. It might help to have approximate measures of uncertainty ([Section 1] promises none). Figure 3 cannot really be interpreted, since positive or negative values may be due to random fluctuations about zero. The contours in Figure 6 are calculated with vastly different precision, and in some respects are incomparable. And, [in Section 6], upon estimating α, β and γ, it would be tempting (but unwise) to assume that such values are significant."

"All random variables in sight are assumed independent. Another way to motivate these weighted models is to assume a multivariate distribution, with the property that the conditional mean at (x,y), given the surrounding data, is a weighted combination of those data. Then the joint distribution exhibits dependence."

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. John Tukey, Rob Tibshirani and the Referee made very helpful comments on the first draft. The research was supported by National Science Foundation Grant DMS-8900613.

APPENDIX I

In this Appendix a few computing details are provided. The census divisions and the province boundaries are specified as polygons. To compute the weights $w_i(x,y)$ an algorithm 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, an algorithm for breaking a polygon up into triangles was required. Such algorithms are discussed in Preparata and Shamos (1985) for example. The approximate 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 and the computations thence carried out on 20 separate work stations. As in Brillinger and Preisler (1983), factors were introduced into the likelihood to stabilize the computations. Miyaoka (1989) found that the computations could be sensitive to the number of nodes employed. In the present series of computations, the number was increased until the results did not change much. There is also the problem of selecting inital values. Here they were taken to be the method of moment estimates, although these are perhaps too inefficient.

APPENDIX II

For simplicity, consider the case of a point process $\{x_j\}$ with rate function ν on the line. The local weighted log likelihood for a Poisson process is, up to a constant,

$$\sum_j W(x - x_j) \log \nu(x_j) - \int W(x - u)\nu(u)du.$$

So, the locally weighted estimate of the rate is

$$\hat{\nu}(x) = \sum_j W(x - x_j) \Big/ \int W(x - u)du,$$

the usual form of estimate. Suppose now the line is broken into intervals R_i, and the aggregate count available is $N(R_i)$. One desires

$$\sum_{x_j \in R_i} W(x - x_j).$$

If this last is to be approximated by $\Theta N(R_i)$, then the method of moments leads to

$$\Theta = \int_{R_i} W(x - u)du \Big/ |R_i|$$

and thence to expression (3).

REFERENCES

ABRAMOWITZ, M., and STEGUN, I.A. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington.

AITCHISON, J., and HO, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.

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). *The Annals of Statistics*, 5, 622-623.

BRILLINGER, D.R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42, 693-734.

BRILLINGER, D.R. (1990). Mapping aggregate birth data. Technical report, Statistics Department, University of California, Berkeley.

BRILLINGER, D.R., and PREISLER, H.K. (1983). Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*, 31-65. New York: Acadamic Press.

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. *Journal of the American Statistical Association*, 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. *Journal of the Royal Statistical Society*, 37, 297-348.

COHEN, A. (1983). Seasonal daily effect on the number of births in Israel. *Applied Statistics*, 32, 228-235.

CRESSIE, N., and READ, T.R.C. (1989). Spatial data analysis of regional counts. *Biometrical Journal*, 31, 699-719.

CROUCH, E.A.C., and SPIEGELMAN, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp\{-t^2\} dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.

DEAN, C., LAWLESS, J.F., and WILLMOT, G.E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics*, 17, 171-182.

DYN, N., and WAHBA, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal on Mathematical Analysis*, 13, 134-152.

FRANKE, R. (1982). Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38, 181-200.

GILCHRIST, W.G. (1967). Methods of estimation involving discounting. *Journal of the Royal Statistical Society*, 29, 355-369.

HINDE, J. (1982). Compound regression models. GLIM 82 (Ed. R. Gilchrist), 109-121. *Lecture Notes in Statistics*, 14. New York: Springer-Verlag.

MALLOWS, C., and TUKEY, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (Eds. Tiago de Oliveira, J. *et al.*), 111-172. London: Academic.

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. *Journal of the American Statistical Association*, 84, 637-650.

MIYAOKA, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. *Canadian Journal Statistics*, 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 Department, Oregon State University.

PREPARATA, F.P., and SHAMOS, I. (1985). *Computational Geometry*. New York: Springer-Verlag.

SHABAN, S.A. (1988). Poisson-lognormal distributions. *Lognormal Distributions*, 195-210 (Eds. E.L. Crow and K. Shimizu). New York: Marcel Dekker.

STANISWALIS, J.G. (1989), The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.

STONE, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595-620.

TIBSHIRANI, R., and HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

TOBLER, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Assocation*, 74, 519-536.

TSUTAKAWA, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 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). Pacific Grove: Wadsworth.

TUKEY, J.W. (1990). Graphical diplays of: Are the (x,y) pairs compatible with a linear dependence? Technical Report No. 301, Princeton University.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 271–277
Statistics Canada

Benchmarking of Economic Time Series

NORMAND LANIEL and KIMBERLEY FYFE[1]

ABSTRACT

Benchmarking is a method of improving 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 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 obtained by applying weighted least squares.

KEY WORDS: Survey errors; Non-linear model; Weighted least squares.

1. INTRODUCTION

Traditionally benchmarking has been defined as the method 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, the monthly shipments of Canadian Manufacturers could be adjusted so that they add up to the Annual Census of Manufacturers shipments figures. Another definition of benchmarking is the more general one 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, the monthly inventories of Canadian Retailers derived from a sample survey could be improved using the end of year inventories obtained from the annual retail trade sample survey. This second definition of benchmarking corresponds to the situation encountered with many economic time series and is the one dealt with in this paper.

The purpose of this article is twofold, first it describes in detail, the benchmarking problem as it appears for many time series produced by large scale economic surveys. Then, two well known benchmarking methods dealing with a single time series are presented and discussed. Since both of these methods fail in some respects to resolve the problem, two other methods which use a non-linear weighted least squares approach are proposed. Finally, two of the above mentioned methods are illustrated with some simulated data and the results are discussed.

2. PROBLEM DESCRIPTION

The problem of improving a two-way table of sub-annual series of estimates with annual series of estimates from business surveys is described here, accompanied by the characteristics of the original data and a list of the features desired from a benchmarking procedure.

[1] Normand Laniel and Kimberley Fyfe, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The sub-annual estimates are often biased due to frame coverage deficiencies. Undercoverage is caused by delay in the inclusion of new businesses and no representation of non-employer businesses (usually small ones) on the frame. These sub-annual estimates are usually derived from relatively small overlapping samples, implying that sampling variances are relatively large and that sampling covariances exist between sub-annual estimates of different time periods. In addition, 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 industry by geographical region tables, where the cells as well as the marginals and the grand totals need to be benchmarked.

As regards annual estimates, they can be assumed to be unbiased since in practice their frames do not suffer much from coverage deficiencies. Also, the annual estimates usually come from relatively large samples or censuses and thus have relatively small or no sampling errors associated with them, while their sampling covariances tend to be large because of substantial sample overlap between years. Another point to note about the annual estimates is that these figures come in approximately two years after the time to which they refer. 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 to which they refer. Therefore, when the sub-annual estimates are to be benchmarked, there will be no annual benchmarks for some of the sub-annual periods.

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 that it can be used without too much data analysis. Second, it must be possible to produce preliminary benchmarking factors for periods for which benchmarks are not yet available. This feature allows benchmarking to be performed as the sub-annual data are produced. Otherwise discontinuities will be introduced in the sub-annual data. It is also desirable that the method maintain consistency between the grand-totals, marginal totals, and cell estimates for the benchmarked estimates in a table.

More discussion on the last two features can be found in Laniel and Fyfe (1989) and (1990) and Cholette (1988a) and (1988b). The rest of this paper deals with the problem of benchmarking a single time series in the context described above.

3. BENCHMARKING A SINGLE SERIES

Four approaches to benchmarking a single time series of sub-annual flow or stock estimates are described in the following sub-sections.

3.1 Denton's Method

In his 1971 paper, Denton proposed procedures for benchmarking based on a Quadratic Minimization approach, each of which corresponds to a specific penalty function. One of these penalty functions is the proportionate first difference between the original and benchmarked series and is often used for the problem of benchmarking time series that was described in section 2. Denton's procedure can be presented in statistical terms by first stating that the sub-annual estimates follow the model:

$$\frac{\theta_t}{y_t} = \frac{\theta_{t-1}}{y_{t-1}} + \epsilon_t, \qquad t=1, 2, \ldots, n \tag{3.1}$$

subject to the restriction to the annual data:

$$z_T = \sum_{t \epsilon T} \theta_t, \qquad T = 1, 2, \ldots, m, \tag{3.2}$$

where:

- t refers to a sub-annual period,
- T refers to an annual period,
- $\{y_t\}$ 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 $(\mathbf{0}, \sigma^2 I)$ and,
- $\{z_T\}$ is a sequence of annual benchmarks.

To find the benchmarked estimates, least squares are applied to the above restricted model.

It is important to note that Denton's approach assumes that the bias follows a random walk and that both the sub-annual and annual data are observed without sampling errors. Unfortunately, these assumptions are unlikely to be satisfied by economic time series (see section 2).

3.2 Hillmer and Trabelsi's Method

In 1987, Hillmer and Trabelsi proposed an alternative approach to benchmarking based on the Box-Jenkins (1976) ARIMA models. They assumed that the sub-annual estimates follow the model:

$$y_t = \theta_t + u_t \qquad t = 1, 2, \ldots, n \tag{3.3}$$

and the annual estimates follow the model:

$$z_T = \sum_{t \epsilon T} \theta_t + a_T \qquad T = 1, 2, \ldots, m, \tag{3.4}$$

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,
- $\{u_t\}$ is a sequence of sub-annual dependent sampling errors with mean vector and covariance matrix $(\mathbf{0}, \Sigma_u)$,
- $\{z_T\}$ is a sequence of annual unbiased estimates, and
- $\{a_T\}$ is a sequence of annual dependent sampling errors with mean vector and covariance matrix $(\mathbf{0}, \Sigma_a)$.

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.

As it is stated with the models, this method takes into account the sampling variances and covariances of the sub-annual and annual estimates. Unfortunately, the approach does not accommodate biases in the sub-annual data. 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) modified the Hillmer and Trabelsi approach by introducing 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 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 estimates follow the model:

$$\frac{y_t}{y_{t-1}} = \frac{\theta_t}{\theta_{t-1}} + v_t \qquad t = 1, 2, \ldots, n \tag{3.5}$$

and the annual estimates follow model (3.4), where:

$\{\theta_t\}$ is a sequence of sub-annual parameters (true values), as in Denton's method,

$\{v_t\}$ is a sequence of dependent sub-annual sampling errors of the trends with mean vector and covariance matrix $(\mathbf{0}, \Sigma_v)$.

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 and the calculation of the covariance matrix of the benchmarked estimates are given in Laniel and Fyfe (1989) or (1990).

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 if the rate of the frame maintenance activities is relatively stable, that is, when the proportion of frame coverage deficiencies is fairly constant over time. This assumption also implies that the undercovered businesses behave like the ones covered by the frame.

One technical problem with this method is that the sampling variance-covariance matrix of the sub-annual 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. For this reason, an alternative model to (3.5) is presented in the next section.

3.4 Model on Levels

The following method is an alternative to the previous one and is suggested so that the sampling variance-covariance matrix of the sub-annual estimates would be easier to obtain. It assumes that the sub-annual estimates follows the model:

$$y_t = \alpha\theta_t + u_t \qquad t = 1, 2, \ldots, n, \tag{3.6}$$

where α is a fixed parameter taking into account the constant relative bias and u_t is the same as for equation (3.3). The annual estimates follow model (3.4).

Benchmarked estimates are obtained by applying least squares theory to the above models. The algorithm required to solve this problem is the same as for method 3.3.

3.5 Discussion

Among the methods reviewed here, the most appropriate one for benchmarking a single time series in the context of the large scale surveys 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, $\hat{\theta}_t$, 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 small number of degrees of freedom also suggests that with the model on levels, we can expect to get benchmarked estimates with a chronological pattern similar to the one observed in the sub-annual data.

A current practical issue with benchmarking methods which take into account sampling errors such as in 3.4, is the derivation of sampling covariances between two level estimates corresponding to two different time periods. Should they be calculated directly using the sample design for all pairs of time periods or should they be modelled? From a theoretical point of view, it is better to calculate these directly, since the sequence of sampling errors is intrinsically a non-stationary stochastic process due to the population variance-covariance varying with time. However, calculating all sampling covariances can be cumbersome, thus leaving the issue of how to obtain sampling covariances still an open question.

3.6 An Example

As a comparison between Denton's method described in section 3.1 and the model on the levels approach suggested in section 3.4, these two methods were applied to a special and interesting benchmarking case. It is a situation where the annual estimates have sampling variances six times the size of the sampling variances of the corresponding monthly estimates. In such a case, the advantage of using the model on levels approach instead of Denton's method will be easily observed.

The special case, though possible in practice, was made up of simulated data. Firstly, twenty-four monthly estimates were taken from an existing economic survey. A sampling covariance matrix was arbitrarily given to these monthly estimates. The variances and covariances were calculated in by using an equal coefficient of variation through time and the following correlation pattern:

$$\rho_{ij} = 1 - \frac{|j - i|}{24} \quad \text{for} \quad i = 1, 2, \ldots, 24 \quad \text{and} \quad j = 1, 2, \ldots, 24$$

where i and j are the indexes of a pair of monthly estimates. Then, two corresponding annual estimates were constructed as follows. The first annual figure was 25% larger than the sum of the first monthly figures. Whereas the second annual figure was only 5% larger than the total of the last twelve monthly observations. The two annual estimates were given sampling variances equal to six times the variances of the corresponding monthly totals and their correlation was fixed at 0.5.

The monthly estimates are represented by the full continuous line and the annual estimates by the horizontal lines on figure 3.1. The two horizontal lines are equal to the values of the

annual figures divided by twelve. On the same figure, the line with long dots represents the monthly series benchmarked with the approach based on the model on levels. The line with short dots is the benchmarked monthly series with Denton's method.

From figure 3.1, it can be observed that the series benchmarked with the model on levels approach has the same year-to-year movement as the original monthly series. Whereas the series benchmarked with Denton's method has the same year-to-year movement as the annual estimates. It can also be seen that both benchmarked series are over the original monthly series.

The difference in the year-to-year movement between the two benchmarked series can be explained as follows. The approach based on the model on levels gives the benchmarked series a year-to-year movement essentially obtained by weighting the annual and sub-annual data with the inverse of their sampling variances. Since, in this example, the sub-annual estimates are much more reliable than the annual estimates, the benchmarked series got the year-to-year movement of the monthly figures. Whereas with Denton's method, the year-to-year movement of the benchmarked series is constrained to one of the annual series regardless of its reliability. In this sense the approach based on the model on levels is better than Denton's method.

As a last comment on this example, the fact that both benchmarked series are above the original monthly series simply illustrates that both methods are providing a correction for the bias of the monthly estimates.

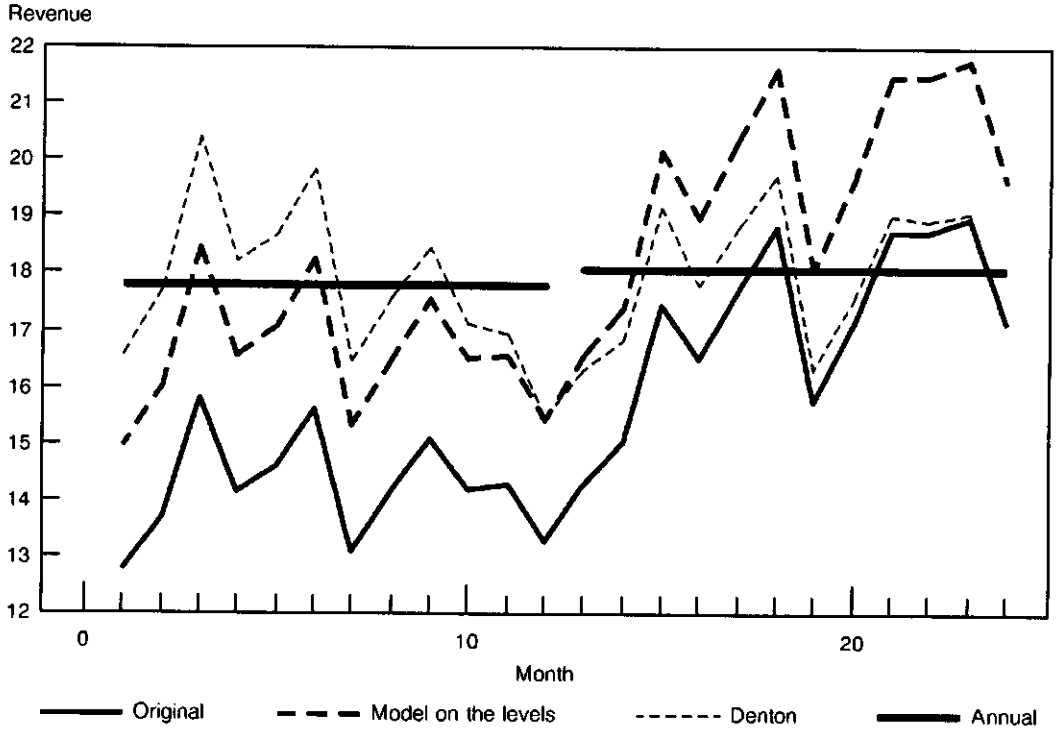


Figure 3.1 Plot of the original and two benchmarked monthly series and of the annual series

4. CONCLUSION

The problem of improving sub-annual survey estimates with the use of 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 to allow its use in an automated mode. The advantage of the procedure over more traditional methods (*e.g.*, Denton's) is that it takes account of sampling errors. Some issues in using the proposed procedure for benchmarking a single time series have been discussed. Two important practical questions have been pointed out: benchmarking a table of series and preliminary benchmarking. Approaches to address these two topics have to be explored.

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.

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*, 66, 99-102.

HILLMER, S.C., and TRABELSI, A. (1987). Benchmarking of Economic Time Series. *Journal of the American Statistical Association*, 82, 1604-1071.

LANIEL, N., and FYFE, K. (1989). Benchmarking of Economic Time Series. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

LANIEL, N., and FYFE, K. (1990). 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*.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 279–282
Statistics Canada

Forgot the Sampling Scheme at the Estimation Stage?

SHIBDAS BANDYOPADHYAY[1]

ABSTRACT

For a class of linear unbiased estimators in a class of sampling schemes, it is shown that one can forget the weights used for sample selection while estimating a population ratio by a ratio of two unbiased estimators, respectively of the numerator and the denominator defining the population ratio. This class of schemes includes commonly used sampling schemes such as unequal probability sampling with or without replacement, stratified proportional allocation sampling with unequal selection probabilities and without replacement in each stratum, *etc.*

KEY WORDS: Ratio of unweighted totals; Symmetric sampling.

1. INTRODUCTION

Let m be the number of adult literates among t adult members in a sample of n families drawn from a given population. Let the population adult literacy rate R be estimated as $r = m/t$. Similarly, for a two- way table giving percentage distribution of persons by age-group and sex, let a cell entry be estimated by a ratio (multiplied by 100 to make it a percentage) of the number of persons classified into the cell to the total number of persons, in the sample of n families.

Irrespective of the method of selection of the families, this simple ratio of two unweighted totals for estimating a ratio or a percentage distribution is acceptable to many non-statistical users. Indeed, in some survey reports, tables giving percentage distributions or rates are so computed, as if the sampling scheme had been a self-weighting one.

If, however, the sampling scheme for selecting the n families had been a (single stage) PPSWOR, one is expected to go about finding weighted totals for obtaining unbiased estimators of numerators and respective denominators before computing a ratio or a percentage distribution.

This study shows that, for sampling schemes such as a single stage PPSWOR but without any further assumptions,

(i) a ratio of two unweighted totals estimates the corresponding population ratio, as a ratio of an *unbiased estimator* of the numerator to an *unbiased estimator* of the respective denominator;

(ii) there is a class of sampling schemes, other than self-weighting designs, for which (i) holds. This class includes one stage unequal probability, with or without replacement, sampling schemes and stratified proportional allocation sampling with unequal probability without replacement selection in each stratum.

2. SYMMETRIC SAMPLING SCHEMES

Consider a finite population consisting of N units $U_1, U_2, \ldots, U_N$. Let Y_i and X_i, denote the values of two study variables, Y and X respectively, associated with the unit U_i, $i = 1, 2, \ldots, N$.

[1] Shibdas Bandyopadhyay; Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

The problem is to estimate a rate or a ratio $R = T(Y)/T(X)$ where $T(Y) = Y_1 + Y_2 + \ldots + Y_N$, and $T(X)$ is similarly defined with the variable X.

The usual procedure is to estimate $T(Y)$ and $T(X)$ unbiasedly and take their ratio to estimate R. The aim of this paper is to follow the same procedure in such a way that the ratio becomes free of the selection probabilities of the sample units.

Fix a sampling scheme.

Let S denote the set consisting of all possible samples such that $p(s) > 0$, where $p(s)$ denotes the probability of drawing the sample s, and $\sum_{s \epsilon S} p(s) = 1$.

For s in S and $i = 1, 2, \ldots, N$,

$n(i,s)$ = the number of times U_i is included in s, and $\alpha_i = \sum_{s \epsilon S} n(i,s)$, the number of times U_i is included in all possible samples.

S, $p(s)$, α_i depend on the sampling scheme.

Definition 2.1. A sampling scheme is said to be symmetric if $\alpha_i = \alpha$, for all $i = 1, 2, \ldots, N$.

The following estimator, based on the sample s, in the class of linear unbiased estimators of Godambe (1955) for $T(Y)$, was studied by Bandyopadhyay *et al.* (1977).

$$T(Y,s) = \sum_{i=1}^{N} Y_i n(i,s)\, \alpha_i^{-1} p^{-1}(s). \tag{2.1}$$

Clearly, $T(Y,s)$ is unbiased for $T(Y)$. An estimator of the ratio $R = T(Y)/T(X)$, as a ratio of an unbiased estimator of $T(Y)$ to an unbiased estimator of $T(X)$, based on a sample s, is

$$R(s) = T(Y,s)/T(X,s) = \sum_{i=1}^{N} Y_i n(i,s)\, \alpha_i^{-1} \Big/ \sum_{i=1}^{N} X_i n(i,s)\, \alpha_i^{-1}. \tag{2.2}$$

For symmetric sampling schemes, $\alpha_i = \alpha$ for all i and (2.2) becomes

$$R(s) = \sum_{i=1}^{N} Y_i n(i,s) \Big/ \sum_{i=1}^{n} X_i n(i,s) = \frac{\text{unweighted total of } Y \text{ values in the sample}}{\text{unweighted total of } X \text{ values in the sample}} \tag{2.3}$$

and the above observations are summarized in the following theorem.

Main theorem. For a symmetric sampling scheme, a ratio of two unweighted totals estimates the corresponding population ratio as a ratio of an unbiased estimator of the numerator to an unbiased estimator of the respective denominator, but the estimated ratio does not involve the selection probabilities of the population units in the sample.

It may be noted that the inclusion probabilities of the units in the sample need not be equal for symmetric sampling schemes. Thus, symmetric sampling schemes need not be self-weighting. Self-weighting designs require constancy of $\alpha_i p(s)$ for all i and s, and constancy of $\alpha_i p(s)$ for all i and s does not make the sampling scheme symmetric.

For a non-symmetric scheme, (2.2) is easy to compute as α_i's are easy to compute in most cases and there is no need to compute inclusion probabilities.

For without replacement sampling of n units, there are $\binom{N-1}{n-1}$ (un-ordered) samples containing a given unit U_i, so $\alpha_i = \binom{N-1}{n-1}$ for all i and thus, in particular, PPSWOR is symmetric. It may be noted that not all PPSWOR schemes result in $\binom{N}{n}$ possible samples. As noted in Connor (1966), in some cases systematic PPS samples in a pre-determined order or randomized PPS systematic sampling may result in zero probability for some set of n units. The result applies if the PPSWOR scheme is such that no joint inclusion probability of any set of n units is zero.

For with replacement sampling of n units, there are N^n (ordered) samples and so $\alpha_i = nN^{n-1}$ for all i and thus, in particular, PPSWR is symmetric.

For PPSWOR in each of k strata, the α-value for each unit in the jth stratum is

$$\alpha_j = \frac{n_j}{N_j} \prod_{i=1}^{K} \binom{N_i}{n_i}$$

which becomes a constant when allocation is proportional and if no joint probability of any set of units in any stratum is zero, where N_j and n_j are respectively the population and sample sizes for the jth stratum, $j = 1, 2, \ldots, k$. Similar allocation may be made to make a multistage sampling scheme symmetric.

For PPSWR sampling, it may be noted that the unbiased estimator of $T(Y)$ given by (2.1) is inadmissible. This estimator can be improved upon by putting $n^*(i,s)$ and α_i^* respectively for $n(i,s)$ and α_i, where $n^*(i,s)$ is 1 if $n(i,s)$ is at least 1 and $n^*(i,s)$ is zero if $n(i,s)$ is zero, and α_i^* is α defined with $n^*(i,s)$. Here, $\alpha_i^* = N^n - (N-1)^n$, the number of (ordered) samples containing a given unit U_i. It has not been possible to obtain a mathematical expression for relative efficiency in a closed form for comparison, even with respect to PPSWR schemes.

Among the possibilitities for comparison of relative bias and relative efficiency, an empirical study is included for comparison with PPSWOR scheme. Another attractive possibility is to study large sample variance and bias using Taylor series expansions.

It is clear that it is not possible to estimate the variance of $R(s)$ without the weights or further assumptions. However, if s_1 and s_2 are two half-samples drawn by the same symmetric sampling scheme (like two independent PPSWOR samples of equal size), R is estimated as $[R(s_1) + R(s_2)]/2$, and its unbiased variance estimator is $[R(s_1) - R(s_2)]^2/4$.

If $T(X)$ is known, a ratio-type estimator for $T(Y)$ is $T(X)T(Y,s)/T(X,s)$, which may be improved as in Bandyopadhyay (1980) depending on whether or not the sampling fraction is more than half.

When the population units are divided into k non-overlapping clusters and the selection probability of the jth cluster is p_j then the design become symmetric with $\alpha_i = 1$ for all units in all the clusters. It may be noted that the sample size is the size of the selected cluster and so, the symmetric sampling schemes need not be fixed sample size designs.

3. EMPIRICAL STUDY ON BIAS AND MEAN SQUARE ERROR

Yates and Grundy (1953) considered the following three hypothetical populations, each with 4 population units.

	Population A				Population B				Population C			
X	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Y	0.5	1.2	2.1	3.2	0.8	1.4	1.8	2.0	0.2	0.6	0.9	0.8

The sampling scheme is to draw a sample of size $n = 2$ by PPSWOR using X-values as size measure. It is proposed to compare bias and mean square error of $R(s)$ with those of $R_{HT}^{(s)}$ where $R_{HT}^{(s)}$ is the ratio of the Horvitz-Thompson (1952) estimator of $T(Y)$ to that of $T(X)$. The result of the comparison is presented below.

Populations:	A	B	C
Relative bias of $R(s)$	0.02456	−0.02785	−0.00496
Relative bias of $R_{HT}(s)$	−0.00379	0.00552	0.00232
MSE of $R(s)$	0.2946	0.2946	0.0824
MSE of $R_{HT}(s)$	0.3159	0.3642	0.0690
Relative efficiency of $R(s)$ to $R_{HT}(s)$	1.0723	1.2362	0.8374

Though the absolute bias of $R(s)$ relative to R is more than that of $R_{HT}^{(s)}$ for the three populations, differences are small. $R(s)$ is a more efficient estimator in populations A and B and $R_{HT}(s)$ is more efficient in population C.

Since the above three populations are more extreme than the situations usually met with in practice, it is anticipated that $R(s)$ may be useful when the sampling scheme is not available at the estimation stage.

ACKNOWLEDGEMENT

The author sincerely appreciates active and constructive comments from the referees leading to the final form of this paper.

REFERENCES

BANDYOPADHYAY, S., CHATTOPADHYAY, A.K., and KUNDU, S.C. (1977). On estimation of population total. *Sankhyā*, Ser. C, 39, 28-42.

BANDYOPADHYAY, S. (1980). Improved ratio and product estimators. *Sankhyā*, Ser. C, 42, 45-49.

CONNOR, W.S. (1966). An exact formula for the probability that two specificed sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-396.

GODAMBE, V.P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, Ser. B, 17, 269-278.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

YATES, F., and GRUNDY, P.M. (1953). Selections without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Ser. B, 15, 253-261.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 283–292
Statistics Canada

Estimation of Panel Correlations for the Canadian Labour Force Survey

HYUNSHIK LEE[1]

ABSTRACT

The Canadian Labour Force Survey uses the rotation panel design. Every month, one sixth of the sample rotates and five sixths remain. Hence, under this rotation scheme, once a rotation panel enters in the sample, it stays 6 months in the sample before it rotates out. Because of this design feature and the way of selecting the rotate-in panel, the estimates based on the panels in the same or different months are correlated. The correlation between two panel estimates is called the panel correlation. Three kinds of panel correlations are defined in this paper: (1) the correlation (denoted by ρ) between estimates for the same characteristic based on the same panel in different months; (2) the correlation (denoted by γ) between estimates of the same characteristic based on geographically neighboring panels in different months; (3) the correlation (denoted by τ) between estimates of different characteristics based on the same panel in the same or different months. This paper describes a methodology for estimating these panel correlations and presents estimated correlations for selected variables using 1980-81 and 1985-87 data with some discussion.

KEY WORDS: Repeated panel survey; Rotation; Taylor method.

1. INTRODUCTION

The Labour Force Survey (LFS) is a continuing monthly household survey which employs rotating panel design. The sample consists of six equal size rotation panels one of which is replaced by a new panel each month. The rotated-in panel stays in the sample for six months before it rotates out from the sample. (For detailed description of the LFS methodology, readers are referred to Platek and Singh (1976) and Singh *et al.* (1990).) Therefore, the estimates based on the same panel consisting of the same sampling units in different months are highly correlated. Moreover, an outgoing rotation panel is usually replaced by a neighboring panel. Because they are geographically close, estimates based on these neighboring rotation panels are also correlated. These correlations are called panel correlations. In this paper, we will describe and discuss how the panel correlations can be estimated and present their estimates for selected variables. The work was originated for the study of composite estimation technique. However, the results are applicable in any situation where the panel correlation plays a role.

The paper is structured as follows. In Section 2, necessary definitions, notations and assumptions are given. Methodology is described in Section 3 and results and discussion are given in Section 4.

2. DEFINITIONS OF PANEL CORRELATION COEFFICIENTS

To define various panel correlations we need to define common panels and the predecessor panel. A panel is identified by the panel number which indicates the duration of the panel in the sample. Thus, Panel 1 in month m, becomes Panel 2 in month $m + 1$, Panel 3 in month $m + 2$,

[1] H. Lee, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Table 1

Common and Predecessor Panels Pertaining to Months m and $m - j$

m	$m-1$	$m-2$	$m-3$	$m-4$	$m-5$	$m-6$	$m-7$	$m-8$	$m-9$	$m-10$	$m-11$
1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))	((2))
2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))
3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))
4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))
5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))
6	5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)

Note: Single and double parentheses indicate *single* and *double* predecessors, respectively.

and so on. Another term *rotation group* is often used to identify a panel regardless of its duration in the sample. For instance, Rotation Group 1 which rotates in in January is identified as Rotation Group 1 throughout its stay in the sample until it rotates out in July. Then, Panel 1 in January indicates Rotation Group 1 and Panel 2 in February indicates the same rotation group which is now two months old and so on.

Two panels in two different months which represent the same rotation group are called *common panels*. When a rotation group rotates out, it is usually replaced by a rotation group consisting of neighboring households and given the same rotation group number. A panel associated with the out-going rotation group is called a *predecessor panel* of a panel associated with the in-coming rotation group. Therefore, in the above example, Panel 6 in June which is associated with Rotation Group 1 is a predecessor panel of Panel 1 in July. Table 1 shows schematically the common and predecessor panels pertaining to given months m and $m - j$.

Since each panel can be identified by two components, month and panel number, let P(month, panel number) denote a panel. Then $\mathrm{P}(m, 4)$ and $\mathrm{P}(m - 1, 3)$, for instance, are common panels 1 month apart. Similarly, $\mathrm{P}(m, 4)$ and $\mathrm{P}(m - 2, 2)$ are common panels 2 month apart. The correlation coefficient of estimates of a characteristic based on common panels that are j months apart is denoted by ρ_j. Obviously, there are no common panels which are more than 5 months apart and thus, the subscript j can be at most 5. We assume that ρ_j is independent of m and panel number. However, it is a function of j and varies between characteristics.

The correlation coefficient of estimates based on a panel and its predecessor that are j months apart is denoted by γ_j. But in this case, j can go up to 11, *i.e.* γ_{11} is the last correlation coefficient in this series and it is the correlation between $\mathrm{P}(m, 6)$ and $\mathrm{P}(m - 11, 1)$. We assume again that γ's are independent of m and panel number. They do, however, depend on characteristic as well as j as ρ-correlations do.

The third type of panel correlation is defined as the correlation between estimates for two different characteristics based on common panels and denoted by τ_j for common panels that are j months apart. Now j can take values from 0 to 5. The same assumptions as for the ρ's and γ's apply here as well.

The formal definitions of ρ's, γ's and τ's are as follows:

Let $y_{m,l}$ be the LFS estimate of a characteristic of interest obtained from $\mathrm{P}(m,l)$. We assume that $V(y_{m,l}) = \sigma_y^2$ regardless of m and l. Then, ρ_j's are defined by

$$\mathrm{Cov}(y_{m,l}, y_{m-j,l-j}) = \rho_j \sigma_y^2, \quad 1 \le j \le 5, \quad j < l \le 6,$$

and γ_j's by

$$\mathrm{Cov}(y_{m,l}, y_{m-j,6+l-j}) = \gamma_j \sigma_y^2,$$

where $1 \le l \le j$ if $1 \le j \le 6$ and $j - 5 \le l \le 6$ if $7 \le j \le 11$.

It would be natural to conjecture that ρ_j's and γ_j's decrease as the subscript j increases and that ρ_j's are larger than γ_j's because ρ_j's are correlations pertaining common households while γ_j's are those pertaining neighboring households. We can also define the correlation between a panel and the predecessor of the panel's predecessor (denoted by double parentheses and called *double predecessor* in Table 1) in a similar way, say δ, and thus, we have $\delta_7, \delta_8, \ldots, \delta_{17}$. They will be smaller than γ_j's but could be quite close to them for the same subscript because double and single predecessors are close geographically. However, the δ-correlations are not considered here due to time and resource constraints.

We assume that $\mathrm{Cov}(y_{m,l}, y_{m,l'}) = 0$ if $l \ne l'$ and $\mathrm{Cov}(y_{m,l}, y_{m-j,l'}) = 0$ if $\mathrm{P}(m-j, l')$ is not a common panel nor a predecessor of $\mathrm{P}(m,l)$.

In order to define τ-correlations, let $x_{m,l}$ be the LFS estimate of another characteristic obtained from $\mathrm{P}(m,l)$ and let $V(x_{m,l}) = \sigma_x^2$ be independent of m and l. Then τ-correlations are defined by

$$\mathrm{Cov}(y_{m,l}, y_{m-j,l-j}) = \tau_j \sigma_x \sigma_y, \quad 0 \le j \le 5, \quad j < l \le 6.$$

3. ESTIMATION OF THE PANEL CORRELATIONS

Since a variance estimation computer program was available, the method described here was geared to use this program with minimum modification. The methodology used in the program is the generalized Keyfitz method (Choudhry and Lee 1987; Lee 1989a) better known as the Taylor method. The program can compute variance estimates of linear combinations of monthly estimates.

We employ the following basic equality to estimate the desired correlations using the existing variance program:

$$\mathrm{Cov}(A,B) = \frac{V(A) + V(B) - V(A-B)}{2}. \tag{1}$$

From the program, $V(A-B)$, $V(A)$ and $V(B)$ can be obtained and so can $\mathrm{Cov}(A,B)$ using (1). An expression for $V(A-B)$ from which (1) can be obtained is also given in Kish (1965).

3.1 Estimation of ρ-Correlations

Let $A = \sum_{l=2}^{6} y_{m,l}$ and $B = \sum_{l=1}^{5} y_{m-1,l}$. A and B are obtained by eliminating Panel 1 from month m and Panel 6 from month $m-1$, respectively. Note that the eliminated panels are uncommon and the remaining ones are all common. Using the variance program we compute estimates of $V(A-B)$, $V(A)$ and $V(B)$ and obtain estimates of $\mathrm{Cov}(A,B)$ by (1). From the assumptions given in Section 2, it is easy to see that

$$\mathrm{Cov}(A,B) = 5\rho_1\sigma_y^2,$$

$$V(A) = V(B) = 5\sigma_y^2,$$

and thus,

$$\rho_1 = \frac{\mathrm{Cov}(A,B)}{\sqrt{V(A)V(B)}}. \tag{2}$$

An estimate of ρ_1 is then obtained by substituting estimates of $\mathrm{Cov}(A,B), V(A)$ and $V(B)$. Estimates of ρ_2, ρ_3 and ρ_4 can be obtained in the same way by putting $A = \sum_{l=j+1}^{6} y_{m,l}$, and $B = \sum_{l=1}^{6-j} y_{m-j,l}, j = 2, 3, 4$. But there is some problem in estimating ρ_5 this way. When we drop all uncommon panels from months m and $m - 5$, only one panel is left in each month and this causes problem in variance estimation for Self-Representing Units (SRUs). SRUs are large cities each of which is represented in the survey by independent sampling. There is no such problem for Non-Self-Representing Units (NSRUs) which are the areas outside of the SRUs, containing rural areas and small urban centers. In NSRUs, each Primary Sampling Unit (PSU), which becomes a replicate for variance estimation, has all rotation panels and thus, even after eliminating 5 uncommon panels, there is still one panel remaining in the PSU so that variance can be computed. In SRUs, however, rotation panels form replicates and if there is only one panel left, then there is only one replicate in each stratum and thus, variance can not be computed in the usual way. Therefore, $\hat{\rho}_5$ was obtained by prediction using a nonlinear regression $\rho = a + bt + ce^{-t}, t = 1, \ldots, 4$. Another way to estimate ρ_5 will be discussed later in Subsection 4.1.

3.2 Estimation of γ-Correlations

It is easy to see that $\mathrm{Cov}(A,B) = (5\rho_1 + \gamma_1)\sigma_y^2$ if $A = \sum_{l=1}^{6} y_{m,l}$ and $B = \sum_{l=1}^{6} y_{m-1,l}$. In general,

$$\mathrm{Cov}(A,B) = \{(6 - j)\rho_j + j\gamma_j\}\sigma_y^2,$$

where

$$A = \sum_{l=1}^{6} y_{m,l},$$

$$B = \sum_{l=1}^{6} y_{m-j,l}, \quad j = 1, \ldots, 4.$$

Then, an estimate of γ_j can be obtained from the following equation:

$$\gamma_j = \frac{1}{j}\left[6\frac{\mathrm{Cov}(A,B)}{\sqrt{V(A)V(B)}} - (6 - j)\rho_j\right], \tag{3}$$

by substituting estimated values on the right. There is a direct way to estimate these γ-correlations including γ_5 by

$$\gamma_j = \frac{\mathrm{Cov}(A_j,B_j)}{\sqrt{V(A_j)V(B_j)}}, \tag{4}$$

where $A_j = \sum_{l=1}^{j} y_{m,l}$ and $B_j = \sum_{l=7-j}^{6} y_{m-j,l}, j = 2, \ldots, 5$. In Section 4, the two methods were compared by using empirical data.

Other γ-correlations (γ_j, $j = 6, \ldots, 10$) are obtained by (4) with

$$A_j = \sum_{l=j-5}^{6} y_{m,l},$$

$$B_j = \sum_{l=1}^{12-j} y_{m-j,l}.$$

There is no simple way of estimating γ_{11} directly or indirectly. Both $\hat{\gamma}_5$ and $\hat{\gamma}_{11}$ were predicted by a log-linear model $\gamma = \exp(a + bt)$, $t = 1, \ldots, 4, 6, \ldots, 10$.

3.3 Estimation of τ-Correlations

These correlations can be estimated by the same way as the ρ-correlations just by replacing $y_{m,l}$ by $x_{m,l}$. Let $A = \sum_{l=j+1}^{6} x_{m,l}$ and $B = \sum_{l=1}^{6-j} y_{m-j,l}$, $j = 0, 1, \ldots, 4$. Then we have

$$\mathrm{Cov}(A,B) = (6 - j)\tau_j \sigma_x \sigma_y,$$

$$V(A) = (6 - j)\sigma_x^2,$$

$$V(B) = (6 - j)\sigma_y^2,$$

from which we get

$$\tau_j = \frac{\mathrm{Cov}(A,B)}{\sqrt{V(A)V(B)}}, \quad j = 0, 1, \ldots, 4. \tag{5}$$

All τ's can be estimated using (5) except τ_5 which is predicted by a log-linear model, $\tau = \exp(a + bt)$, $t = 1, \ldots, 4$.

4. RESULTS AND DISCUSSION

By using the methods discussed in the previous section, estimates of ρ- and γ-correlations were computed from the 1980-81 and 1985-87 LFS data for 5 characteristics: In Labour Force (IN LF), Employed (EMP), Employed Agriculture (EMP AG), Employed Non-Agriculture (EMP NON-AG), Unemployed (UNEMP). The panel correlations were estimated for only 3 provinces, Nova Scotia (NS), Ontario (ONT), and British Columbia (BC) from the 1980-81 data. However, the estimation was extended to all provinces when more recent data (March 1985 - February 1987) were used. Moreover, 4 more characteristics, the employed and the unemployed of two age groups, 15-24 and 25+ (EMP 15-24, EMP 25+, UNEMP 15-24, UNEMP 25+), were added. The estimation of τ-correlations was done only for those additional characteristics for NS, ONT and Alberta (ALT) from the 1985-87 data.

In the following, only part of these results will be presented and discussed. All the results are available in Lee (1989b).

4.1 Estimates of ρ-Correlations

The results of estimated ρ-correlations are given in Table 2. Even though estimates for the 5 characteristics (IN LF, EMP, EMP AG, EMP NON-AG, UNEMP) from the 1985-87 data are available for all provinces, the results for only 3 provinces, NS, ONT and BC, are presented for a historical comparison. Table 2 also shows the results for the other 4 characteristics (EMP 15-24, EMP 25+, UNEMP 15-24, UNEMP 25+) from the provinces of NS and ONT.

The ρ-correlations are generally high as expected because they are correlations for the common panels. The correlations for EMP AG are the highest and those for UNEMP are the lowest. It seems that the size of the ρ-correlation indicates the degree of mobility of the labour force with a particular characteristic. For instance, the high ρ-correlation for EMP AG shows a low mobility of the labour force in agriculture while a high mobility of unemployed labour force is demonstrated in its low ρ-correlation. The different levels of mobility of labour force in two age groups are also evident. The younger group (15-24) is more mobile than the older one (25+).

The decreasing trend of the ρ-correlations over time is clearly demonstrated in the results. The trend was extremely well fitted by a nonlinear regression model $\rho_t = a + bt + ce^{-t}$. The R-squares (multiple correlations) are close to 1 (> 0.98). Therefore, the predicted values for ρ_5 seem to be very good. In Lee (1989a and 1989b), $\hat{\rho}_5$ was obtained by extrapolating $\hat{\rho}_3$ and $\hat{\rho}_4$ instead. The differences between the predicted and extrapolated values for $\hat{\rho}_5$, however, are very small. They are less than 0.01 for all characteristics except for UNEMP, UNEMP 15-24 and UNEMP 25+ where the largest difference is 0.03.

Table 2

Estimates of ρ-Correlations (1980-81 and 1985-87 Data)

Prov	Characteristic	80–81 Data					85–87 Data				
		$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
NS	IN LF	0.862	0.797	0.744	0.679	0.622	0.845	0.769	0.730	0.696	0.670
	EMP	0.866	0.783	0.714	0.651	0.590	0.863	0.768	0.713	0.686	0.660
	EMP AG	0.913	0.837	0.756	0.678	0.598	0.912	0.867	0.825	0.802	0.773
	EMP NON-AG	0.865	0.774	0.710	0.649	0.594	0.873	0.779	0.724	0.697	0.670
	UNEMP	0.590	0.455	0.333	0.243	0.145	0.703	0.546	0.426	0.415	0.375
	EMP 15-24						0.773	0.632	0.556	0.495	0.446
	EMP 25+						0.878	0.800	0.754	0.729	0.705
	UNEMP 15-24						0.618	0.454	0.364	0.300	0.246
	UNEMP 25+						0.695	0.554	0.443	0.440	0.406
ONT	IN LF	0.843	0.782	0.717	0.674	0.622	0.846	0.781	0.732	0.681	0.635
	EMP	0.852	0.779	0.709	0.664	0.611	0.853	0.771	0.706	0.648	0.592
	EMP AG	0.955	0.926	0.901	0.861	0.827	0.962	0.948	0.944	0.937	0.934
	EMP NON-AG	0.861	0.791	0.724	0.678	0.625	0.866	0.795	0.746	0.701	0.660
	UNEMP	0.580	0.445	0.334	0.286	0.222	0.579	0.436	0.328	0.291	0.238
	EMP 15-24						0.747	0.605	0.500	0.429	0.356
	EMP 25+						0.888	0.824	0.777	0.732	0.691
	UNEMP 15-24						0.468	0.339	0.257	0.219	0.178
	UNEMP 25+						0.622	0.468	0.365	0.313	0.256
BC	IN LF	0.849	0.767	0.705	0.665	0.622	0.817	0.753	0.701	0.647	0.597
	EMP	0.835	0.755	0.695	0.651	0.607	0.851	0.770	0.711	0.651	0.597
	EMP AG	0.896	0.809	0.733	0.656	0.582	0.938	0.886	0.847	0.828	0.805
	EMP NON-AG	0.855	0.769	0.715	0.661	0.616	0.857	0.784	0.730	0.679	0.632
	UNEMP	0.516	0.407	0.334	0.320	0.294	0.634	0.524	0.459	0.363	0.290

4.2 Estimates of γ-Correlations

As mentioned in Subsection 3.2, there are two ways of estimating γ_2, γ_3 and γ_4, that is, by formulae (3) and (4). We will call the method by (3) as Method 1 and that by (4) as Method 2. Only Method 1 can be used to estimate γ_1 while direct estimation of γ_5 is feasible only by Method 2. The two methods are compared in Table 3 using empirical data. In the table, $\hat{\gamma}_5$'s for Method 1 are predicted values by a log-linear model. The table shows that the two methods produced somewhat different results. The correlations produced by Method 2 clearly show an increasing trend contrary to our intuition while Method 1 gave more acceptable results. Moreover, if we compare these correlations with $\hat{\gamma}_1$ in Table 4A (which had to be estimated by Method 1), Method 1 seems to produce more reasonable results than Method 2. Therefore, we adopted Method 1. However, if everything is correct, the two methods should be equivalent and produce similar results. It seems that the real data do not conform to some extent with the assumptions we made to derive the formulae.

Estimates of the γ-correlations are presented in Tables 4A and 4B. The size of γ-correlations is much smaller than that of ρ-correlations as we expected. But it also reflects differences in mobility of the labour force with different characteristics as seen from the results of ρ-correlations.

The overall trend of $\hat{\gamma}$'s is somewhat fuzzy, especially for the results from the 1985-87 data. There are about 25% of cases – a case is a row entry in the tables – in Table 4B which show an increasing trend. In those cases, the log-linear regression lines have a positive slope even though it is fairly small in magnitude. Moreover, in most of those cases, R-squares are small, which indicates that fittings by the log-linear model are not good. This does not mean, however, that there are other models which can fit the data better. Rather it means that no clear trend is exhibited. Among the cases that show a decreasing trend, about half of the cases have an R-square greater than 0.5.

The results from the 1980-81 data show a quite different picture. There is only one case that shows an increasing trend and most of the cases have R-squares > 0.5. In fact, the results for NS and BC look more reasonable than those for ONT as far as the trend is concerned.

Table 3

Comparison of Estimates of γ_2, γ_3, γ_4 and γ_5 Obtained by Different Methods (Ontario, 1980-81)

Characteristic	Method	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$
IN LF	1	0.141	0.128	0.133	0.135
	2	0.107	0.105	0.116	0.120
EMP	1	0.136	0.142	0.142	0.147
	2	0.100	0.115	0.126	0.133
EMP AG	1	0.483	0.474	0.486	0.451
	2	0.321	0.370	0.407	0.448
EMP NON-AG	1	0.150	0.147	0.157	0.163
	2	0.117	0.134	0.145	0.149
UNEMP	1	0.074	0.076	0.063	0.080
	2	0.043	0.056	0.046	0.043

Note: Methods 1 and 2 are defined by the formulae (3) and (4) in Section 3, respectively.

Table 4A

Estimates of γ-Correlations
(1980-81 Data)

Prov	Characteristic	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
NS	IN LF	0.288	0.263	0.265	0.250	0.236	0.233	0.211	0.199	0.193	0.167	0.164
	EMP	0.262	0.219	0.228	0.226	0.219	0.239	0.210	0.200	0.188	0.161	0.172
	EMP AG	0.351	0.308	0.283	0.237	0.205	0.190	0.141	0.113	0.063	0.021	0.007
	EMP NON-AG	0.238	0.187	0.189	0.180	0.164	0.151	0.123	0.121	0.136	0.091	0.086
	UNEMP	0.106	0.176	0.091	0.097	0.091	0.076	0.066	0.063	0.066	0.032	0.031
ONT	IN LF	0.161	0.141	0.128	0.133	0.135	0.136	0.125	0.127	0.124	0.122	0.117
	EMP	0.164	0.136	0.142	0.142	0.147	0.149	0.148	0.150	0.153	0.141	0.146
	EMP AG	0.477	0.483	0.474	0.486	0.451	0.474	0.459	0.429	0.394	0.323	0.368
	EMP NON-AG	0.184	0.150	0.147	0.157	0.163	0.167	0.166	0.169	0.174	0.156	0.165
	UNEMP	0.141	0.074	0.076	0.063	0.080	0.051	0.045	0.060	0.077	0.136	0.074
BC	IN LF	0.177	0.137	0.117	0.119	0.119	0.112	0.101	0.112	0.094	0.066	0.070
	EMP	0.211	0.146	0.133	0.107	0.101	0.083	0.050	0.068	0.058	−0.033	−0.015
	EMP AG	0.380	0.311	0.301	0.272	0.241	0.216	0.198	0.170	0.122	0.078	0.071
	EMP NON-AG	0.207	0.166	0.161	0.129	0.108	0.093	0.069	0.038	0.023	−0.004	−0.020
	UNEMP	0.126	0.125	0.114	0.103	0.091	0.076	0.062	0.092	0.032	0.040	0.031

Table 4B

Estimates of γ-Correlations
(1985-87 Data)

Prov	Characteristic	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
NS	IN LF	0.250	0.238	0.247	0.230	0.216	0.204	0.181	0.196	0.189	0.162	0.160
	EMP	0.170	0.183	0.205	0.196	0.185	0.157	0.158	0.194	0.198	0.219	0.198
	EMP AG	0.326	0.296	0.246	0.245	0.265	0.267	0.234	0.217	0.259	0.269	0.231
	EMP NON-AG	0.146	0.168	0.199	0.201	0.178	0.153	0.152	0.189	0.199	0.216	0.201
	UNEMP	0.233	0.267	0.241	0.211	0.206	0.168	0.171	0.176	0.157	0.187	0.147
	EMP 15-24	0.107	0.127	0.140	0.133	0.112	0.105	0.099	0.107	0.090	0.074	0.082
	EMP 25+	0.088	0.075	0.117	0.108	0.100	0.099	0.090	0.103	0.099	0.137	0.118
	UNEMP 15-24	0.051	0.080	0.042	0.024	0.054	0.061	0.079	0.081	0.058	0.011	0.049
	UNEMP 25+	0.155	0.129	0.177	0.171	0.148	0.159	0.158	0.127	0.102	0.134	0.124
ONT	IN LF	0.162	0.138	0.141	0.134	0.132	0.135	0.127	0.116	0.111	0.103	0.101
	EMP	0.114	0.122	0.121	0.122	0.117	0.124	0.119	0.108	0.110	0.112	0.111
	EMP AG	0.508	0.518	0.553	0.561	0.571	0.569	0.582	0.617	0.668	0.650	0.672
	EMP NON-AG	0.133	0.140	0.132	0.140	0.157	0.156	0.168	0.182	0.204	0.205	0.210
	UNEMP	0.030	0.047	0.055	0.047	0.043	0.048	0.039	0.030	0.039	0.048	0.041
	EMP 15-24	0.012	−0.006	0.018	0.031	0.017	0.023	0.011	0.011	0.016	0.044	0.029
	EMP 25+	0.354	0.358	0.349	0.343	0.319	0.312	0.298	0.285	0.276	0.240	0.246
	UNEMP 15-24	0.068	0.039	0.038	0.058	0.033	0.026	0.008	0.018	0.011	−0.002	−0.006
	UNEMP 25+	0.052	0.054	0.033	0.017	0.034	0.033	0.026	0.018	0.021	0.044	0.022
BC	IN LF	0.103	0.095	0.113	0.103	0.090	0.090	0.091	0.083	0.078	0.030	0.055
	EMP	0.125	0.100	0.112	0.111	0.116	0.135	0.123	0.121	0.118	0.095	0.114
	EMP AG	0.394	0.443	0.426	0.401	0.396	0.400	0.401	0.381	0.347	0.334	0.345
	EMP NON-AG	0.080	0.067	0.076	0.072	0.091	0.109	0.111	0.118	0.112	0.106	0.124
	UNEMP	0.096	0.086	0.084	0.080	0.083	0.097	0.068	0.074	0.068	0.083	0.071

Table 5

Estimates of τ-Correlations
x_1: EMP 15-24, x_2: EMP 25+, x_3: UNEMP 15-24, x_4: UNEMP 25+,
(1985-87 Data)

Province	Characteristic	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
NS	(x_1,x_2)	0.150	0.140	0.148	0.181	0.187	0.196
	(x_1,x_3)	−0.440	−0.275	−0.187	−0.135	−0.039	0.126
	(x_1,x_4)	−0.036	−0.040	−0.043	−0.015	0.024	0.022
	(x_2,x_3)	−0.029	−0.037	−0.078	−0.049	−0.016	−0.038
	(x_2,x_4)	−0.437	−0.374	−0.276	−0.182	−0.231	−0.094
	(x_3,x_4)	0.136	0.127	0.094	0.055	0.049	0.020
ONT	(x_1,x_2)	0.092	0.070	0.055	0.040	0.028	0.010
	(x_1,x_3)	−0.420	−0.267	−0.205	−0.161	−0.145	−0.010
	(x_1,x_4)	−0.065	−0.056	−0.053	−0.036	−0.028	−0.019
	(x_2,x_3)	−0.061	−0.054	−0.054	−0.042	−0.089	−0.074
	(x_2,x_4)	−0.392	−0.303	−0.230	−0.187	−0.181	−0.077
	(x_3,x_4)	0.058	0.043	0.022	0.013	0.022	0.001

4.3 Estimates of τ-Correlations

Table 5 contains estimates of τ-correlations obtained from the 1985-87 data for all possible combinations of EMP 15-24 (denoted by x_1), EMP 25+ (x_2), UNEMP 15-24 (x_3) and UNEMP 25+ (x_4). The correlations between x_1 and x_2 are positive as well as those between x_3 and x_4. Other correlations are mostly negative. In terms of magnitude, only the correlations pertaining to (x_1, x_3) and (x_2, x_4) are quite different from zero. Others are close to zero. These observations seem to agree with what we understand about the movement of labour force between the employed and the unemployed in the same age group. When the employment increases, the unemployment decreases and vice versa. The trend is obviously upward in these cases.

The data were fit by a log-linear model and τ_5's were predicted. The model fitting seems reasonable except for the correlations between (x_2, x_3) whose R-squares are very small in both provinces NS and ONT.

4.4 Conclusions

The estimation of correlations from complex survey data is a difficult problem. It is so not because the derivation of formulae is difficult - in fact, the formulae given here are elementary - but because there are many practical constraints in applying the formulae. If we had not made the assumptions in Section 3, the estimation of the panel correlations by using the existing computer program would have been impossible. On the other hand, these assumptions should be conformable to the real data to which the formulae are applied. In our case, there seem to be some unconformable elements in the assumptions we made to the real data, which was indicated by the discrepancy in the results obtained by formulae (3) and (4) (see Table 3). Nevertheless, the estimates are not thought to be unreasonable.

In a study of the composite estimator for the LFS, the results given in this paper were successfully used to compare various composite estimators (Kumar and Lee 1983). Recently

Binder and Dick (1990) proposed a method for analyzing Seasonal ARIMA models by taking the survey errors into account. They applied their technique to the LFS data using the estimated panel correlations. However, in cases when the results to be obtained by the use of the estimated panel correlations are sensitive to the accuracy of these estimates, the results should be interpreted carefully.

ACKNOWLEDGEMENT

The author would like to thank the anonymous referees, Editor M.P. Singh, and Assistant Editor L. Mach for their helpful comments. The earlier version of this paper was benefited by comments of S. Kumar and Y. Bélanger of Statistics Canada.

REFERENCES

CHOUDHRY, G.H., and LEE, H. (1987). Variance estimation for the Canadian Labour Force Survey. *Survey Methodology*, 13, 147-161.

BINDER, D.A., and DICK, J.P. (1990). Analysis of seasonal ARIMA models from survey data. *Survey Methodology*, this issue.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.

LEE, H. (1989a). Variance estimation methodology and general purpose variance estimation system for the Labour Force Survey. Methodology Working Paper Series, SSMD-89-022 E, Statistics Canada.

LEE, H. (1989b). Estimation of panel correlations for the Canadian Labour Force Survey. Methodology Working Paper Series, SSMD-89-023 E, Statistics Canada.

PLATEK, R., and SINGH, M.P. (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

SINGH, M.P., DREW, J.D., GAMBINO, J., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 293–304
Statistics Canada

First Wave Effects in the U.S. Consumer Expenditure Interview Survey

ADRIANA R. SILBERSTEIN[1]

ABSTRACT

Panel responses to the U.S. Consumer Expenditure Interview Survey are compared, to assess the magnitude of telescoping in the unbounded first wave. Analysis of selected expense categories confirms other studies' findings that telescoping can be considerable in unbounded interviews and tends to vary by type of expense. In addition, estimates from the first wave are found to be greater than estimates derived from subsequent waves, even after telescoping effects are deducted, and much of these effects can be attributed to the shorter recall period in the first wave of this survey.

KEY WORDS: Bounding; Telescoping; Recall Bias; Conditioning.

1. INTRODUCTION

Respondents to retrospective surveys are asked to recall details of events within a specific time interval, or reference period, and this task of identifying the correct time in which events occurred may be as difficult as remembering the events. Misdating, or "telescoping", is widely recognized as a source of error in surveys, although it is rarely studied directly (Neter and Waksberg 1965). Respondents tend to include in the report events that occurred outside the reference period (external telescoping), *e.g.*, when events are recalled as more recent than they actually are (forward telescoping). Data that can be validated with independent records show that both forward and backward misdating errors are made by respondents (Mathiowetz 1985). This could be "due to the respondent's wish to perform the task required.... When in doubt, the respondent prefers to give too much information rather than too little" (Sudman and Bradburn 1974, p. 69). The net effect of telescoping is generally forward. Bounding methods are designed to create boundaries around the reference period of the survey report, and, in so doing, avoid misdating errors by respondents. A method for bounding the starting point of the reference period, best applied during the interview, involves comparing events reported in a prior interview and deleting duplicate reports. Extending the reference period up to the interview day is a method commonly used to bound the end of the reference period. "Unbounded" reports result by necessity from one-time surveys, and for questions asked only once or for the first time in panel surveys, since no prior data exist to check for erroneous inclusions. These effects can be reduced by including "anchoring" techniques during the interview, *e.g.* constructing a time line (Mingay 1987, p. 132).

This paper is concerned with reporting levels experienced by first time respondents of panel surveys, and provides a comparative analysis of first and subsequent interview waves. The study investigates potential telescoping, conditioning, and recall length effects in estimates of household expenditures, based on data reported in the U.S. Consumer Expenditure (CE) Interview Survey for the year 1984. This survey is one of two independent components designed to collect national data on household expenditures, the other component being the Diary Survey.

[1] Adriana R. Silberstein is a Mathematical Statistician, Office of Prices and Living Conditions, Statistical Methods Division, U.S. Bureau of Labor Statistics, Washington, D.C. 20212, USA.

The survey is conducted by the Census Bureau under contract to the Bureau of Labor Statistics. The first wave of the CE Interview Survey is used to establish cooperation, collect initial inventory data on household possessions, and bound the second wave. There are four subsequent waves of interviews three months apart, collecting data for the previous three calendar months up to the interview day. The bounding method is as follows. Expenses reported for the portion of the calendar month in which the interview takes place (or "current month") are later transcribed onto the next wave questionnaire; this information is available to the interviewer to check for duplicate reports, but is not read to respondents. Data collected during the first wave pertain to expenditures for the current month and for one previous calendar month; these latter expenditures are excluded entirely from the estimates, while current month expenditures become part of the second wave. More details on collection and estimation methods can be found in the 1984 Bulletin (U.S. Bureau of Labor Statistics 1986), and are discussed by Silberstein and Jacobs (1989).

The findings underscore the need for bounding methods in retrospective data collection, since sizable telescoping effects may be present in unbounded recall. In addition, the analysis points out that first time responses may yield higher estimates even after telescoping effects are deducted. These first wave effects may be a direct result of the shorter recall in this wave of the CE Interview Survey, although other factors are not excluded. A discussion of the analysis used to identify telescoping effects is included in section 2, and estimates of telescoping and first wave effects are included in section 3. Conclusions can be found in section 4.

2. IDENTIFYING TELESCOPING EFFECTS

2.1 Method of Analysis

One approach for identifying telescoping errors, discussed by Kalton *et al.* (1989, p. 257), is to examine whether there are duplicates in individual responses to consecutive waves. This micro-level approach is not necessarily accurate, as the respondent for a given household may change from one wave to the next. The method is also impractical, since independent records, needed to reconcile discrepancies on dates, may not be readily available. Duplicate responses may not be recorded as such in an ongoing survey, even when they are identified during the interview, as in the CE Interview Survey. More commonly, telescoping effects are evaluated at the aggregate level, by comparing estimates of unbounded and bounded responses, with certain precautions. Tracking the experience of several panels is advisable in order to overcome seasonal incomparabilities, since bounded responses are reported subsequently to unbounded responses and, therefore, do not refer to the same time interval. Another factor to account for in the comparisons is panel conditioning, a phenomenon that refers to changes in respondent behavior as a result of being part of a panel, or to changes in the quality of reports. The assumptions made and the method of estimation used in this study are discussed in section 3, whereas the preliminary testing procedure is described here.

The first step in the analysis is to ascertain whether symptoms of external telescoping can be detected from the survey data. A level of reporting in the first wave that is higher than expected is an indication of telescoping. Unbounded interviews are known to yield higher estimates than bounded interviews, as documented in several studies that compared unbounded and bounded responses (Neter and Waksberg 1964 and 1965; Murphy and Cowan 1976; Cantor 1985). Another indication is the presence of differential effects across separate types of the collected data. Major sources of differences in the way events are retrieved and stated by respondents are recall bias and telescoping. The relationship of these factors suggests that

smaller expenses are forgotten as time increases, but larger more salient expenses, that tend to be remembered better, are more often telescoped.

Telescoping errors can also occur in bounded responses, causing the forward shifting of data within the reference period (internal telescoping). While overall estimates do not change as a result of these effects, the distribution for the three recall months is affected. Reports of apparel and home furnishing and equipment expenses were selected for the study, because characteristics of these expenses were helpful in the analysis. These commodities include expenditures of various degree of salience, and were grouped accordingly. They also tend to differ by degree of underreporting. Many apparel estimates are 40% below the estimates from the National Accounts (NA), and several estimates for home furnishings and equipment are also lower than NA estimates. Estimates for furniture and selected equipment categories, on the other hand, are only 7% below the independent estimates (Gieseman 1987, p. 11), and higher reports in the first wave can be interpreted as the result of external telescoping.

The hypothesis evaluated is whether the first recall month of bounded waves, *i.e.*, the month prior to the interview, is reported similarly to the past month in the first wave. The Hotelling T^2 was used to test differences in eight expenditure groups within each of the two commodities. Given two vectors of means in a repeated-measures design, a two-tailed .05-level test of H_0: $C\mu = 0$ (equality of means) versus H_1: $C\mu \neq 0$ was applied. H_0 was rejected if:

$$[(C\bar{x})'(CSC')^{-1}C\bar{x}]/[np/(n-(p-1))] > F_{p,n-p+1}(.05), \qquad (1)$$

where $\bar{x}$ is a vector of sample means within each commodity (ordered as shown in the tables), S is the covariance matrix computed with the method of balanced repeated replication ($n = 20$ replicates), C is the contrast matrix shown below, and p is the number of contrasts in C.

$$\underset{(px2p)}{C} = \left[\begin{array}{cccc|cccc} 1 & 0 & .. & 0 & -1 & 0 & .. & 0 \\ 0 & 1 & .. & 0 & 0 & -1 & .. & 0 \\ . & . & .. & . & . & . & .. & . \\ . & . & .. & . & . & . & .. & . \\ 0 & 0 & .. & 1 & 0 & 0 & .. & -1 \end{array}\right].$$

Simultaneous confidence intervals for individual comparisons by group were derived using the Bonferroni method (Johnson and Wichern 1988), with percentile $t_n(.05/2p)$. Expenditure means were computed using a log transformation of individual expenses reported in the first recall month. Sample weights included adjustments for nonresponse and subsampling, but excluded final weight factors for population controls, which were not available for the first wave. Note that weight adjustments for the first wave were computed only as part of this research, since they are not needed in the ongoing estimation process.

Data from waves 2 to 5 were combined, since differences between these waves were very small. Responses by participants in all five waves (3200 respondents) were selected to assure comparability between the waves and bounding of waves 2 to 5. Unbounded interviews are experienced by new panel respondents, *e.g.* new occupants at a sample address, and by respondents who do not participate in one or more wave during the panel. In 1984, 89% of the interviews in waves 2 to 5 were bounded, 8% were unbounded because respondents were new to the panel and 3% were unbounded resulting from a previous refusal or other noncooperation (Silberstein 1988). Estimates are affected by unbounded responses, as pointed out by Biderman and Cantor (1984), but this aspect is not treated directly in this study.

2.2 Test Results

Comparisons between means are shown in Table 1 in the original scale, *i.e.*, without the log transformation used in the statistical tests. The first wave displays higher means in nearly all expense groups, and the overall test is significant. The tests for the individual groups reveal that significant differences are found only for large expenditures, such as coats and jackets in apparel and appliances and furniture in home furnishings and equipment. The groups with significant differences are more represented in wave 1 than in other waves, not surprisingly: they account for 19% of total apparel and 72% of total home furnishings in the first wave, compared to 16% and 67%, respectively, in the first recall month of other waves, as shown in Table 2 (columns 1 and 2). A greater number of expenses are also reported in wave 1 for these groups of expenses (Table 2, columns 3 and 4). In addition, the average dollar value of reported expenses in wave 1 tends to be different from the other waves for big-ticket items (*e.g.*, major appliances), but very similar for smaller items (Table 2, columns 5 and 6).

Table 1

Percent Difference in Expenditure Means

	Wave 1 Versus First Recall Month of Waves 2 to 5	
	% Difference (a)	*s*
APPAREL: (b)	**14.5***	**4.9**
Coats, jackets, furs, suits	39.6*	12.9
Trousers, slacks, jeans	13.6	9.5
Shirts, blouses, tops	9.7	5.6
Sweaters, dresses, skirts	16.4	4.7
Undergarments, hosiery	6.9	5.4
Miscellaneous and combined clothing	−2.5	7.3
Footwear	2.1	6.1
Other apparel items and services	27.4	25.4
Overall test value: 4.16*		
HOME FURNISHINGS AND EQUIPMENT: (b)	**48.6***	**8.4**
Major appliances	76.1*	27.5
Other appliances	56.3*	17.0
Furniture	111.0*	24.8
Large household and entertainment equipment	34.2*	16.0
Other household and entertainment equipment	19.1*	7.1
Home furnishing repair and services	7.0	14.6
Dishes, decorative items, linens	14.0	16.0
Floor and window coverings	52.5	24.3
Overall test value: 13.86*		

(a) Positive values indicate first wave mean is greater. Base of percentages is mean of first recall month in waves 2 to 5.
(b) Commodity totals not included in overall test.
s Standard error of percent difference.
* Significant (α = .05).

Table 2

Comparisons of First Wave and First Recall Month of Subsequent Waves

	Percent of Total Expenses		Percent of Total Number of Expenses		Average Dollar Value of Expenses	
	Wave 1	Waves 2 to 5	Wave 1	Waves 2 to 5	Wave 1	Waves 2 to 5
	(1)	(2)	(3)	(4)	(5)	(6)
APPAREL:	**100.0**	**100.0**	**100.0**	**100.0**	**$ 35**	**$ 33**
Coats, jackets, furs, suits	19.2	15.7	9.3	8.6	71	59
Trousers, slacks, jeans	10.7	10.8	10.6	9.8	36	35
Shirts, blouses, tops	10.0	10.4	12.0	12.2	31	29
Sweaters, dresses, skirts	14.3	14.0	13.0	12.4	38	37
Undergarments, hosiery	5.2	5.6	16.8	16.7	11	11
Miscellaneous and combined clothing	15.5	18.2	15.4	16.4	36	38
Footwear	11.7	13.1	12.8	13.6	33	31
Other items and services	13.5	12.2	10.1	10.4	45	40
HOME FURNISHINGS AND EQUIPMENT:	**100.0**	**100.0**	**100.0**	**100.0**	**$123**	**$ 92**
Major appliances	11.4	9.6	4.2	3.4	370	277
Other appliances	2.3	2.2	9.2	7.1	29	30
Furniture	28.3	19.9	8.9	7.5	385	251
Large household and entertainment equipment	19.7	21.8	8.8	7.6	262	266
Other household and entertainment equipment	10.7	13.4	22.7	22.8	58	56
Home furnishing repair and services	4.7	6.6	8.4	9.5	67	65
Dishes, decorative items, linens	12.9	16.8	33.1	37.5	46	39
Floor and window coverings	10.0	9.8	4.6	4.5	294	172

These differences can be interpreted in several ways, *e.g.*, they may indicate that more expensive purchases are reported in the first wave, or that purchases reported in the first wave are remembered as more expensive. Another interpretation is that a period of time longer than a month may be covered by respondents when the recall is unbounded, especially for large, easily remembered, expenses. In Table 3, comparisons by wave are extended to include the three recall months of subsequent waves. The findings are consistent with the previous tests, but tend to narrow in on the issue of telescoping effects. These comparisons are made on the basis of reporting rates according to the dollar value of the expense. The reporting rate is defined as the percentage of respondents reporting one or more expense of a given type. Note that individual expenses are generally entered on the questionnaire, with the exception of expenses for the same item, month and person in the family, which are usually reported as combined totals and counted as one "expense".

Table 3

Monthly Reporting Rates by Expense Size

	Wave 1	Waves 2 to 5 by Recall Month		
		First	Second	Third
		Percent of respondents		
	(1)	(2)	(3)	(4)
APPAREL:				
No Apparel Expenses (a)	28.8	29.3	38.2	45.5
Less than $10	38.4	37.7	27.9	25.4
$ 10 to $ 40	57.9	55.2	45.3	41.0
$ 40 to $100	35.1*	31.0	26.5	21.0
$100 and over	17.0*	13.7	11.5	8.8
Wave 1 vs 1st recall month of waves 2 to 5 **Overall test value: 29.1***				
HOME FURNISHINGS AND EQUIPMENT:				
No Home Furnishing Expenses (a)	48.1*	51.2	58.5	62.4
Less than $10	12.3	12.5	7.5	7.5
$ 10 to $ 40	30.9	30.0	25.0	22.1
$ 40 to $100	21.3*	18.4	14.9	12.8
$100 to $400	18.7*	13.8	12.1	10.3
$400 and over	8.6*	5.6	5.1	4.6
Wave 1 vs 1st recall month of waves 2 to 5 **Overall test value: 17.0***				

(a) Category included in overall test.
* Significant (α = .05).

Consistent with the previous comparisons, the overall test is significant and the individual comparisons show that significantly more respondents report expenses of $100 or more in the first wave; reporting rates for smaller expenses are not significantly different, instead. When the three recall months are examined, the reporting rates for the first recall month appear to be closer to the first wave than to the other two months. The three recall months in waves 2 to 5 show a familiar pattern of decreased reporting, and noteworthy is the increase in the percent of respondents reporting "no expenses". This pattern is evident in each panel wave, as documented by Silberstein and Jacobs (1989) and further studied by Silberstein (1989), and is more likely due to recall effects than telescoping. When reporting rates are recomputed to include only respondents that report the commodity, it is found there are more similarities among the three recall months in subsequent waves than with the first wave. (The rates can be derived from Table 3, by using the percentage of reporters with expenses as the base.) These reporting rates for home furnishing items of $100 and over are 53% in the first wave and 40%, 41%, and 40%, respectively, in the three recall months of other waves. For apparel items of $100 and over the rates are 24% in the first wave and 19%, 19%, and 16%, respectively, in the three recall months of other waves. These differences are believed to be symptomatic of external telescoping in the unbounded recall.

3. ESTIMATING TELESCOPING AND FIRST WAVE EFFECTS

3.1 Telescoping Effects

The hypothesis of equality of means implied the response task in the first wave is similar to the one experienced for the first recall month in subsequent waves. The data did not support the hypothesis, since differential effects were found, suggesting external telescoping in the first wave. The results tend to agree with the notion, forwarded by Loftus (1986, p. 196), that internal telescoping may "arise from a different cognitive mechanism" than external telescoping. A general definition of external telescoping (β), on a monthly basis and assuming no panel conditioning, is given by the ratio of unbounded one month recall (with sample mean $\bar{x}_U$) and bounded one month recall (with sample mean $\bar{x}_B$):

$$\beta = (E\bar{x}_U / E\bar{x}_B) - 1. \tag{2}$$

This expression may be an overstatement since conditioning effects contribute to lower values for the bounded mean. Panel responses commonly display a downward trend, due to decreased reporting with increasing time-in-sample (TIS) (Bailar 1989). Conditioning effects (α) between two consecutive waves can be defined by the ratio of the two responses (with sample means $\bar{x}_i$ and $\bar{x}_{i+1}$):

$$\alpha = 1 - (E\bar{x}_{i+1} / E\bar{x}_i). \tag{3}$$

A number of assumptions were made to develop telescoping estimates from the survey data. Expenditure means of bounded one month recall, needed for comparisons with the first wave, cannot be obtained directly from the three month recall. Monthly means computed by dividing the bounded three month recall by a factor of three are not acceptable, considering the recall loss evident in the third recall month of the CE Interview Survey. As an alternative, the first and second recall months were used to estimate bounded monthly means, assuming that recall bias in the second month is moderate and telescoping into the first recall month is mostly from the second recall month. The estimating method is an adaptation of the model developed by Neter and Waksberg in analyzing the 1960 experimental study of expenditures for Residential Alterations and Repairs (Neter and Waksberg 1964 and 1965). The model implies that telescoping and conditioning effects are multiplicative and conditioning compounds with time-in-sample. Since conditioning effects are derived from relationships observed between second and third waves, two terms are necessary when estimating (2) under the assumption of conditioning. An estimate of telescoping is therefore:

$$b_C = (\bar{x}_U / \bar{x}_B)(1 - a)(1 - a/2) - 1. \tag{4}$$

The derivation of (4) is given in the appendix. The conditioning rate (a) was assumed to be constant between waves, considering the special subset of respondents in all five waves. (The Neter/Waksberg model assumed greater effects between the first and second wave.) Time-in-sample effects appear to be small in the CE Interview Survey, judging from a study that compared responses in waves 2 to 5 (Silberstein and Jacobs 1989). An explanation for this may be that declines in reporting are offset by improvements in reporting, as respondents become more knowledgeable about the reporting process. Two conditioning assumptions provided two estimates of telescoping effects, using (4): $a = 0$ (no conditioning), and $a > 0$ conditioning, equal to the rate observed between second and third waves. Four apparel groups and three home furnishing and equipment groups showed some decline from second to third waves, displayed as positive proportions in column 5 of Table 4. These ratios, while not

Table 4

Telescoping Estimates Based on Expenses

	Telescoping effects b_c				TIS effects
	If $a = 0$		If $a > 0$		a
	%	s	%	s	
	(1)	(2)	(3)	(4)	(5)
APPAREL:	**28.4**	**7.0**	-	-	**−0.02**
Coats, jackets, furs, suits	46.2	14.2	-	-	−0.01
Trousers, slacks, jeans	30.3	8.6	12.3	11.8	0.10
Shirts, blouses, tops	27.7	7.8	17.6	16.7	0.05
Sweaters, dresses, skirts	28.3	5.9	8.7	15.0	0.11
Undergarments, hosiery	22.2	6.9	7.2	12.7	0.08
Miscellaneous and combined clothing	5.2	9.5	-	-	−0.18
Footwear	18.1	7.1	-	-	−0.08
Other items and services	54.9	35.8	-	-	−0.15
HOME FURNISHINGS AND EQUIPMENT:	**63.1**	**8.9**	-	-	**−0.04**
Major appliances	95.4	30.7	-	-	−0.03
Other appliances	76.4	16.1	36.0	19.7	0.16
Furniture	113.3	25.2	-	-	−0.05
Large household and entertainment equipment	38.7	13.1	36.5	33.7	0.01
Other household and entertainment equipment	26.2	8.9	-	-	−0.11
Home furnishing repair and services	15.6	14.5	-	-	−0.29
Dishes, decorative items, linens	45.4	14.4	-	-	−0.06
Floor and window coverings	89.4	38.0	66.8	68.7	0.08

a Time-in-sample (TIS), or conditioning, effects when positive.
s Standard error of percent difference.

significant (.05 level), were applied as the conditioning loss between the first and the second wave. Net increases in reports were not considered realistic for the unknown conditioning between these two waves.

The results give indications of the increase that would occur in the estimates in the absence of bounding. Table 4 shows estimates of telescoping effects in percentage form, excluding conditioning effects (column 1), and including them (column 3). Telescoping levels of 40% or higher are estimated for "Coats, *etc.*" and "Other items and services" (a group that includes watches and jewelry), but much lower levels are estimated for other apparel groups. High telescoping levels (63%, on average) are estimated for home furnishing and equipment expenses. Telescoping estimates decrease considerably when some conditioning effects are taken into account, and would be even lower if greater conditioning effects were assumed between wave 1 and wave 2. While these estimates are affected by sampling variability and the assumptions made, the results are consistent with findings reported in other surveys. Neter and Waksberg (1965) reported average telescoping effects of 55% with no conditioning losses and 39% with conditioning losses, for home improvement expenditures; telescoping effects were much lower for small jobs. Telescoping effects derived from the 1974/75 Crime Survey indicated telescoping effects of 44% for personal victimization incidents and 40% for property victimization (Murphy *et al.* 1976).

3.2 First Wave Effects

Differences in responses between first and subsequent waves reflect many cognitive aspects of panel interviews. This section discusses some of the factors involved, and includes a preliminary investigation of net effects. Provided that respondents participate in the whole panel, there is a progressive relationship between respondent and interviewer and more clear expectations on both sides. Quite a few interview conditions change, however. While in some panel surveys subsequent waves may be presented as follow-ups to the first wave, in the CE Interview Survey respondents are asked to report for a period of time three times as long after the first wave and detailed income information is asked in waves 2 and 5. This greater reporting load, and a resulting faster interview pace, has a negative impact on reporting levels, even for the first recall month of these waves. More expense records, *e.g.*, check books and bills, may be used in these waves compared to the first wave, making the bounded reports less likely to be affected by telescoping within the three recall months. The first wave is an easier interview, especially with regard to categories of expenses sensitive to the length of the reference period and the number of persons in the household, *e.g.* apparel expenses. The relative importance of these factors should be researched in field and laboratory studies.

Separate estimates of first wave means, net of telescoping, were developed using the two sets of telescoping effects shown in Table 4. These means ($\bar{x}_{B1}$) were derived by dividing the unbounded means by the telescoping estimates:

$$\bar{x}_{B1} = \bar{x}_U/(1 + b_C). \tag{5}$$

Results are summarized by commodity in Table 5. Both estimates of net first wave means are higher than means of waves 2 to 5 for all recall months combined, shown in column 2. The total apparel mean is 10% higher in the first wave when conditioning effects are not included, and 16% higher when they are included. The home furnishing and equipment means are also higher, but at a smaller scale: 3% without conditioning and 5% with conditioning. These estimated effects, remaining after telescoping, are interpreted as resulting from the shorter recall period and lesser reporting load in the first wave. The differences between the two commodities and the results for specific groups of expenditures imply that potential gains in reporting tend to increase for smaller expenses, but become quite marginal for big-ticket items.

Table 5

Summary Comparisons of FirstWave and Subsequent Waves
Annual Expenditure Means (Standard errors)

	Wave 1	Waves 2 to 5 All Recall Months (a)	Waves 2 to 5 First recall Month	Wave 1 Net of Telescoping	
				Assuming no TIS Effects	Assuming TIS Effects
	(1)	(2)	(3)	(4)	(5)
APPAREL	$1,663 (59.6)	$1,182 (61.7)	$1,452 (71.0)	$1,295 (66.2)	$1,370 (n.a.)
HOME FURNISHINGS AND EQUIPMENT	$1,972 (85.0)	$1,179 (59.7)	$1,327 (73.1)	$1,209 (61.5)	$1,235 (n.a.)

(a) Means differ from published 1984 estimates, due to special subset of respondents and missing final weight factors.

4. CONCLUSIONS

This paper provides an investigation of potential telescoping effects in unbounded interviews. These effects appear to be considerable, especially for more salient or prominent events. Results from the U.S. Consumer Expenditure Interview Survey indicate that estimates of large infrequent expenses, based on unbounded one month recall, may be between 30% and 50% overstated. Lower overstatement levels are more likely in estimates of small frequent expenses. These findings are in close agreement with other studies on the subject. The study demonstrates that external telescoping effects are much greater than internal telescoping effects within a three month recall period of subsequent waves. In addition, the first wave of the panel survey studied was found to exhibit higher means than the overall means for subsequent waves, even after estimated telescoping effects were deducted. Since the first wave in this survey has one month recall, it is concluded that considerable improvements in reporting levels can be expected from a shorter recall. The potential gains are estimated to be at least 10% for frequent expenditures, but would become marginal as the value of the expenditure increases.

Although the one month recall is viewed as the major reason for the higher estimates, other factors are not excluded. Conditioning effects, assumed constant in this study, may vary between waves. Estimates of one month recall would be even greater, if higher conditioning effects were assumed between the first and second waves. Cognitive aspects of the interview, *e.g.*, respondents cooperation and involvement, and interviewers' approach to collecting data, should be researched in order to understand panel conditioning. The issue of differential effects by type of expenditure should also be addressed within this context. Field and laboratory studies of these data collection aspects would have implications for improving panel survey methodology.

ACKNOWLEDGMENTS

The author thanks the referees, Stuart Scott, and Sylvia Leaver for their comments.

APPENDIX

(1) Explanation of Selected Expenditure Groups

SELECTED APPAREL

Miscellaneous and combined clothing: nightwear, loungewear, accessories, uniforms, and clothing items for infants under 2.

Other apparel items and services: watches, jewelry, sewing materials for making clothes, repair and alteration services, and clothing rental or storage.

SELECTED HOME FURNISHINGS AND EQUIPMENT

Other appliances: small electric kitchen and personal care appliances.

Large household and entertainment equipment: lawn mowers, window air conditioners, televisions, sound equipment, and bicycles.

Other household and entertainment equipment: radios, tape recorders, tools, calculators, camping or sports equipment, and infants equipment.

(2) Estimates of Telescoping Effects

(Adapted from: Neter and Waksberg (1965), 33-37).

For each expenditure group

Let: $\bar{x}_U$ = unbounded one month recall sample mean;

$\bar{x}_B$ = bounded one month recall sample mean, not directly observed in the CE Interview Survey;

$\bar{x}_2, \bar{x}_3$ = one-month-average sample means from waves 2 and 3, respectively, computed using first and second recall months.

Define: Telescoping effect β, assuming no conditioning

$$\beta = (E\bar{x}_U/E\bar{x}_B) - 1. \tag{1}$$

Conditioning effect, α, between two consecutive waves

$$\alpha = 1 - (E\bar{x}_{i+1}/E\bar{x}_i). \tag{2}$$

Then, assuming telescoping compounds on conditioning,

$$\beta_C = (E\bar{x}_U/E\bar{x}_B)(1 - \alpha) - 1 \tag{3}$$

is the telescoping effect under conditioning.

Using the estimated conditioning effect between 2nd and 3rd waves, $a = 1 - (\bar{x}_3/\bar{x}_2)$, the estimated mean for bounded one month recall is:

$$\begin{aligned} \bar{x}_B &= (\bar{x}_2 + \bar{x}_3)/2 \\ &= (\bar{x}_2 + \bar{x}_2(1 - a))/2 \\ &= \bar{x}_2(1 - a/2). \end{aligned} \tag{4}$$

Assuming a constant rate of conditioning and using (3) and (4), an estimate of the telescoping effect under conditioning, b_C, is:

$$b_C = (\bar{x}_U/\bar{x}_B)(1 - a)(1 - a/2) - 1. \tag{5}$$

REFERENCES

BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 1-24.

BIDERMAN, A.D., and CANTOR, D. (1984). A longitudinal analysis of bounding, respondent conditioning, and mobility as sources of panel bias in the national crime survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 708-713.

CANTOR, D. (1985). Operational and substantive differences in changing the NCS reference period. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.

GIESEMAN, R. (1987). The consumer expenditure survey: Quality control by comparative analysis. *Monthly Labor Review*, March, 8-14.

JOHNSON, R.A., and WICHERN, D.W. (1988). *Applied Multivariate Statistical Analysis*. 2nd Edition, Englewood Cliffs, New Jersey: Prentice Hall, 188-190.

KALTON, G., KASPRZYK, D., and McMILLEN, D. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 249-270.

LOFTUS, E. (1986). Survey remembering. *Proceedings of the Second Annual Research Conference*, Bureau of the Census, Washington, D.C., 193-207.

MATHIOWETZ, N.A. (1985). The problem of omissions and telescoping error: New evidence from a study of unemployment. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 482-487.

MINGAY, D.J. (1987). Report on the consumer expenditure survey. *Questionnaire Design: Report on the 1987 BLS Advisory Conference*, (Eds. J. Bienias, C. Dippo, and M. Palmisano). Bureau of Labor Statistics, Washington, D.C., 129-138.

MURPHY, L.R., and COWAN, C.D. (1976). Effects of bounding on telescoping in the National Crime Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 633-638.

NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

NETER, J., and WAKSBERG, J. (1965). Response errors in collection of experimental data by household interviews: An experimental study. Technical Report No. 11, Bureau of the Census, Washington, D.C.

SILBERSTEIN, A.R. (1988). Selected first-interview effects in the Consumer Expenditure Interview Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 485-490.

SILBERSTEIN, A.R., and JACOBS, C.A. (1989). Symptoms of repeated interview effects in the Consumer Expenditure Interview Survey. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 289-303.

SILBERSTEIN, A.R. (1989). Recall effects in the U.S. Consumer Expenditure Interview Survey. *Journal of Official Statistics*, 2, 125-142.

SUDMAN, S., and BRADBURN N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

U.S. BUREAU OF LABOR STATISTICS (1986). *Consumer Expenditure Survey: Interview Survey, 1984*. Washington, D.C., Bulletin No. 2267.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 305–330
Statistics Canada

Symmetry in Flows Among Reported Victimization Classifications with Nonresponse

ELIZABETH A. STASNY[1]

ABSTRACT

The United States' National Crime Survey is a large-scale, household survey used to provide estimates of victimizations. The National Crime Survey uses a rotating panel design under which sampled housing units are maintained in the sample for three-and-one-half years with residents of the housing units being interviewed every six months. Nonresponse is a serious problem in longitudinal data from the National Crime Survey since as few as 25% of all individuals interviewed for the survey are respondents over an entire three-and-one-half-year period. In addition, the nonresponse typically does not occur at random with respect to victimization status. This paper presents models for gross flows among two types of victimization reporting classifications: number of victimizations and seriousness of victimization. The models allow for random or nonrandom nonresponse mechanisms, and allow the probabilities underlying the gross flows to be either unconstrained or symmetric. The models are fit, using maximum likelihood estimation, to the data from the National Crime Survey.

KEY WORDS: Categorical data; Ignorable nonresponse; Longitudinal survey; National Crime Survey; Nonignorable nonresponse.

1. INTRODUCTION

The United States' National Crime Survey (NCS) is a large-scale, household survey conducted by the U.S. Bureau of the Census for the Bureau of Justice Statistics. Data from the NCS is used to produce quarterly estimates of victimization rates and yearly estimates of the prevalence of crime. The survey uses a rotating panel of housing units (HU's) under which individuals living in sampled HU's are interviewed up to seven times at six-month intervals.

Individuals interviewed for the NCS are asked about crimes committed against them or against their property in the previous six months. In this work, we begin to explore the victimization status reported by households (HH's) within sampled HU's from one interview to the next. Victimization status for a HH will be considered in two ways: by the number of crimes reported (zero, one, and two or more) and by the type of crime reported (no crime, property crime, and personal contact crime).

Since responses are not available from one NCS interview period to the next for all HH's, we must decide how to handle missing observations. The nonresponse problem is a serious problem in the longitudinal data available from the NCS. For example, Fienberg (1980) noted that complete, three-and-one-half-year records of NCS interviews are available for as few as 25% of all individuals interviewed. In addition, the nonresponse typically does not occur at random with respect to victimization status (see, for example, Saphire (1984)).

This work extends the models developed by Stasny (1986) for nonrandom nonresponse in estimating gross flows. In particular, the models presented here allow for symmetry in the matrix of flows among victimization classifications as well as allowing for completely random nonresponse, ignorable nonrandom nonresponse, or nonignorable nonresponse.

[1] Elizabeth A. Stasny, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.

Section 2 of this paper provides a brief description of the NCS and the longitudinal data from the survey. Section 3 gives a general form of the models for symmetry in gross flow matrices with missing data and presents iterative procedures for obtaining maximum likelihood estimators (MLE's) for the parameters of the models. Section 4 describes the fits of the models to data from the NCS. Section 5 presents conclusions and suggests areas for future research.

2. THE NATIONAL CRIME SURVEY AND DATA

2.1 Survey Design

The NCS is a stratified, multi-stage, cluster sample of HU's. The survey was begun in July 1972 by the Law Enforcement Assistance Administration but has been administered by the Bureau of Justice Statistics since December 1979. The target population for the NCS is the civilian, non-institutionalized population of persons aged 12 and over living in housing units. The survey provides information on personal and household crimes committed against the individuals in sampled HU's. The following crimes and attempted crimes are covered by the NCS: assault, auto or motor vehicle theft, burglary, larceny, rape, and robbery. Crimes not covered by the survey include kidnapping, murder, shoplifting, and crimes that occur at places of business.

The NCS uses a rotating panel design under which a sampled HU is maintained in the sample for three and one-half years with interviews conducted at six-month intervals for a total of seven possible interviews. The initial interview at each HU, however, serves as a bounding interview and is not used for the purpose of estimation. Although there is a six-month interval between interviews at any one HU, NCS interviews are conducted in every month of the year; in order to make efficient use of trained interviewers, one-sixth of the HU's in the sample are scheduled for interviews each month. Since the sampling unit for the NCS is the HU, no attempt is made to follow individuals who move away from the HU during the three-and-one-half-year period. Rather, new individuals entering the HU are included in the survey. Each different group of individuals who live in a HU during its time in the NCS sample is considered a separate HH.

NCS interviews are conducted for all individuals 12 years of age or older who live in the sampled HU at the time of the interview. During the interview, individuals are asked about crimes committed against them or against the household in the previous six months. A single HH respondent is asked a series of six screening questions to elicit information on crimes committed against the HH (burglary, larceny, and motor vehicle theft). Then an eleven-question screener is used to elicit information from each individual in the HH concerning personal crimes committed against that individual (assault, rape, and robbery). An incident report is completed for each crime mentioned in response to the screening questions.

Additional information on the design and history of the NCS is provided, for example, by the U.S. Department of Justice and Bureau of Justice Statistics (1981), Saphire (1984), Dodge and Skogan (1987), and Montagliani (1987). A new sample design for the NCS has been used since January 1986. Taylor (1987) describes the redesign of the NCS and research associated with the redesign effort. The data used in this work, however, were collected under the original NCS design.

2.2 The Longitudinal Data

The data used in this work are from a large, longitudinal data set which includes all the regular NCS interview information collected from January 1975 to June 1979 except for the HU's that rotated into the sample in 1979. To make it easier to handle the data, this research uses only a subset of the data. The subset was created by taking a random start at the record

for the eighth HU in the full data set and then every fifteenth record after that. The resulting data set contains NCS records for 12,432 HU's. Because the HU's on the original longitudinal file are ordered in such a way that units from the same cluster appear together, the 1-in-15 systematic sample should not include two or more HU's from a single cluster. Thus, this research does not consider the problem of correlations among HU's within clusters.

2.3 Flows Among Victimization Classifications

The hierarchical, longitudinal data were used to create summary matrices for the years 1975, 1976, 1977, and 1978 showing flows among reported victimization classifications from each HH's first interview in a year to the HH's second interview for the year. Note that, since NCS interviews are conducted every month of the year, the first interview may occur at any time from January through June and the second interview may occur in July through December. Depending on the month of the interview, the victimizations reported in the first interview are those that occurred between the previous July and May while those reported in the second interview occurred between January and November. Thus, the analysis here explores only the reporting of crimes from one interview to the next. It cannot, for example, address issues of change in victimization reporting at various times of the year except in a very general sense.

It should be noted that during the time when the data were collected, a reference-period experiment was conducted using a sample of NCS HU's. Since individuals in HU's included in the experiment were asked to report victimizations for reference periods other than the usual six-month period, those HU's were not used in this analysis.

For the analyses here, each HH interviewed at least once during a given year was classified according to its reporting and victimization status at the two interview times. A victimization may have been reported by any member of the HH and may be against an individual or against the HH. Two sets of matrices showing victimization classifications are used in the analyses of Section 4. The matrices are given in Appendix I.

The first set of matrices show cross-classifications of HH's by the number of victimizations reported in the first and second interviews for each year. The classifications are: crime free (no victimizations reported), single crime (one victimization reported), multiple crime (two or more victimizations reported), and missing (HH did not respond or rotated out of the sample). The second set of matrices show cross-classifications of HH's by the type of victimization reported. The classifications are: crime free, property crime (burglary, larceny, and motor vehicle theft), contact crime (rape, assault, robbery, purse snatching, and pocket picking), and missing. These type-of-crime groupings are the same as those used in the NCS. In cases where multiple crimes were reported by a single HH, the classification used is for the most serious crime reported (contact crimes are taken to be more serious than property crimes).

Notice the large amount of nonresponse in the observed matrices shown in Appendix I. Only about 50% of the HH's who responded in at least one of the two interviews responded at both interview periods. The models presented in the following section, will allow us to handle this nonresponse while exploring the structure of the underlying matrix of probabilities of flows among the victimization classifications.

3. THE MODELS

This section presents a general form of the models that will be used to explore gross flows among victimization classifications in the NCS data. The form of the models follows that proposed by Chen and Fienberg (1974) for contingency tables with completely and partially classified data. The models for nonresponse are those developed by Stasny (1986) as well as

a model for random nonresponse. The model for symmetry in the flows, however, does not appear in the previous work. The models are presented in a general form because they are applicable to problems other than estimating gross flows among victimization classifications using NCS data.

3.1 Model for the Observed Data

Consider observation units that respond to a survey in at least one of two interview periods. Suppose that, when a unit responds to the survey, that unit is classified into one of K classifications. If a unit does not respond to the survey, that unit is classified as missing. Then the interview-to-interview flow data may be represented as in Table 1.

Table 1
Summary of Observed Data

		Time 2: 1	2	...	K	Missing
Time 1	1	x_{11}	x_{12}	...	x_{1K}	x_{1M}
	2	x_{21}	x_{22}	...	x_{2K}	x_{2M}
	⋮	⋮	⋮	⋮	⋮	⋮
	K	x_{K1}	x_{K2}	...	x_{KK}	x_{KM}
	Missing	x_{M1}	x_{M2}	...	x_{MK}	,

where x_{ij} = number of units with survey or missing status i at time 1 and j at time 2.

We suppose that each unit would fall into one of the cells of the $K \times K$ matrix of survey classifications if it were observed at both interview times. Let p_{ij} be the probability that a unit has status i at time 1 and status j at time 2, where i and j take on the values 1, 2, ..., K. Each unit in the (i,j) cell of the matrix of survey classifications has a chance of being missing at one of the two survey times. Let λ_{tij} be the probability that a unit in the (i,j) cell loses its classification at time t and, hence, is classified as missing at that time. Then the probabilities underlying the observed data are as shown in Table 2.

Table 2
Probabilities Underlying Observed Data

		Time 2: 1	2	...	K	Missing
Time 1	1	$\{(1-\lambda_{1ij}-\lambda_{2ij})p_{ij}\}$				$\left\{\sum_{j=1}^{K} p_{ij}\lambda_{2ij}\right\}$
	2					
	⋮					
	K					
	Missing	$\left\{\sum_{i=1}^{K} p_{ij}\lambda_{1ij}\right\}$				.

Assuming that the p_{ij} are probabilities from a multinomial distribution, the likelihood function for the observed data is proportional to

$$\left\{\prod_{i=1}^{K}\prod_{j=1}^{K}\left[p_{ij}(1-\lambda_{1ij}-\lambda_{2ij})\right]^{x_{ij}}\right\}$$
$$\times\left\{\prod_{i=1}^{K}\left[\sum_{j=1}^{K}p_{ij}\lambda_{2ij}\right]^{x_{iM}}\right\}$$
$$\times\left\{\prod_{j=1}^{K}\left[\sum_{i=1}^{K}p_{ij}\lambda_{1ij}\right]^{x_{Mj}}\right\}.$$

There are $3K^2 + 2K - 1$ free parameters defined above and only $K^2 + 2K$ observed cells of data with a single constraint on the total sample size. Thus there are too many parameters to estimate using the observed data and we must reduce the number of parameters in the model. In the following we reduce the number of parameters to be estimated by considering two models for the p_{ij}-parameters and six models for the λ_{tij}-parameters.

3.2 Models for the p and λ Probabilities

We consider two models for the p_{ij}'s, the probabilities of flows among survey classifications: the unconstrained model and the model of symmetric flows. Under the model of unconstrained flow probabilities, there is a different probability, p_{ij}, for every (i,j) cell of the flow matrix. Under the model of symmetric flows, we have $p_{ij} = p_{ji}$ for $i \neq j$ so that the probability that a unit has survey classification i at time 1 and j at time 2 is the same as the probability that a unit has survey classification j at time 1 and i at time 2. Note that symmetry in the cell probabilities of the flow matrix implies equality of row and column marginal totals. Thus the model of symmetry in flow probabilities implies a certain stability in the population since the expected number of units with a particular survey classification at time 1 is the same as the number with that classification at time 2.

As defined above, the λ_{tij}'s, the probabilities that units with survey classifications i at time 1 and j at time 2 are missing at time t, depend on the time at which the nonresponse occurs and on the survey classifications at both times 1 and 2. We consider six simpler models for these probabilities. These models, along with the associated degrees of freedom under both models for the p_{ij}, are given below:

	d.f. unconstrained p_{ij}	d.f. symmetric p_{ij}
Model R: $\lambda_{tij} = \lambda$,	$2K - 1$	$(K^2 + 3K - 2)/2$
Model A: $\lambda_{1ij} = \lambda_{1j}$, $\lambda_{2ij} = \lambda_{2i}$,	0	$(K^2 - K)/2$
Model B: $\lambda_{tij} = \lambda_t$,	$2K - 2$	$(K^2 + 3K - 4)/2$
Model C: $\lambda_{1ij} = \lambda_j$, $\lambda_{2ij} = \lambda_i$,	K	$(K^2 + K)/2$
Model D: $\lambda_{1ij} = \lambda_{1i}$, $\lambda_{2ij} = \lambda_{2j}$,	0	$(K^2 - K)/2$
Model E: $\lambda_{1ij} = \lambda_i$, $\lambda_{2ij} = \lambda_j$,	K	$(K^2 + K)/2$

Model R is the model of random nonresponse. Under Model R, there is a single probability of nonresponse for all units at both times regardless of survey classification. Under Model A, the probability that a unit is missing at time t depends on both the time and the survey classification at the time when the unit responds. Note that if Model A is used for the λ-parameters and the unconstrained model is used for the p_{ij}, then the model is a saturated model which will fit the data exactly. Under Model B, the probability that a unit is missing at time t depends only on the time. Under Model C, the probability that a unit is missing at time t depends only on the unit's survey classification at the time when the unit responds. Under Model D, the probability that a unit is missing at time t depends on both the time and the survey classification at the time when the unit is missing. If Model D is used for the λ-parameters and the unconstrained model is used for the p_{ij}, then the model is a saturated model which will fit the data exactly. Under Model E, the probability that a unit is missing at time t depends only on the unit's survey classification at the time when the unit is missing.

Under Model R, nonresponse is said to be completely at random. Under Models A, B, and C, nonresponse is said to be ignorable nonresponse in that the nonresponse mechanism depends only on the observed data. Nonresponse under Models D and E is nonignorable nonresponse since the nonresponse mechanism depends on the missing data. (See Little and Rubin (1987) for more information on the types of nonresponse.)

In the following two subsections, we describe procedures for fitting the models presented above. The fits of the models can be assessed using either the Pearson X^2 statistic or G^2, the likelihood ratio statistic. Both statistics have asymptotic χ^2 distributions, with degrees of freedom as shown above, given that the model is correct. In the following we use the notation "Model R-U" to denote the pairing of Model R for the λ-parameters and the unconstrained model for the p_{ij}. "Model R-S" will denote the pairing of Model R for the λ-parameters and the symmetric model for the p_{ij}. Similar notation will be used to denote the pairings of Models A, B, C, D, and E for the λ-parameters with one of the two models for the p_{ij}.

3.3 Estimation of the p and λ Parameters Under Models R, A, B, and C

The likelihood functions for the eight models created using one of the two models for the p_{ij} and Model R, A, B, or C for the λ_{tij} factor into two pieces: one piece a function of the p-parameters alone and one a function of the λ-parameters alone. Thus, the MLE's may be found separately for the two sets of parameters. In addition, the p-parameter estimates do not depend on which of these four models is used for the λ-parameters, and the λ-parameter estimates do not depend on which of the two models is used for the p-parameters.

An iterative procedure for obtaining MLE's for the p-parameters under the unconstrained model paired with Model R, A, B, or C for the λ-parameters is given in Chen and Fienberg (1974). The equations for this procedure are provided in Appendix II.

Under the symmetric model for the p-parameters paired with Model R, A, B, or C for the λ-parameters, the factor of the likelihood equation involving only the p_{ij}'s is as follows:

$$\left\{\prod_{i=1}^{k} p_{ii}^{x_{ii}}\right\} \times \left\{\prod_{i=1}^{k}\prod_{j=i+1}^{k} p_{ij}^{x_{ij}}\right\} \times \left\{\prod_{i=2}^{k}\prod_{j=1}^{i-1} p_{ji}^{x_{ji}}\right\}$$

$$\times \left\{\prod_{i=1}^{k} p_{i\cdot}^{x_{iM}}\right\} \times \left\{\prod_{j=1}^{k} p_{j\cdot}^{x_{Mj}}\right\}, \qquad (1)$$

where a dot in a subscript indicates summation over that subscript. Equation (1) is maximized subject to the constraint that the sum of the p_{ij}'s is one. In general, an iterative procedure is required to obtain the MLE's. Let $x_{..} = \sum_{i=1}^{K}\sum_{j=1}^{K} x_{ij}$ be the total number of units observed at both times and let $n = x_{..} + x_{.M} + x_{M.}$ be the total number of units observed in at least one of the two interview times. Then the iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Symmetric p_{ij} Under Models R, A, B, and C

1. $p_{ii}^{(0)} = x_{ii}/x_{..}$

   $p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..}$ for $i \neq j$.

2. $p_{ii}^{(\nu+1)} = \left[x_{ii} + (x_{iM} + x_{Mi})p_{ii}^{(\nu)}/p_{i.}^{(\nu)}\right]/n$

   $p_{ij}^{(\nu+1)} = \left[(x_{ij} + x_{ji}) + (x_{iM} + x_{Mi})p_{ij}^{(\nu)}/p_{i.}^{(\nu)} + (x_{jM} + x_{Mj})p_{ij}^{(\nu)}/p_{j.}^{(\nu)}\right]/2n$ for $i \neq j$.

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other positive values satisfying the constraint that the p_{ij}'s sum to one may be used.

An iterative procedure for obtaining MLE's for the λ-parameters under Model A and the closed-form estimator for the λ-parameters under Model B are given in Chen and Fienberg (1974). An iterative procedure for obtaining MLE's for the λ-parameters under Model C is given in Stasny (1986). The equations for these procedures are provided in Appendix II.

Under Model R for the λ-parameters, the factor of the likelihood equation involving only λ is as follows:

$$\left\{\prod_{i=1}^{K}\prod_{j=1}^{K}(1-2\lambda)^{x_{ij}}\right\} \times \left\{\prod_{i=1}^{K}\lambda^{x_{iM}}\right\} \times \left\{\prod_{j=1}^{K}\lambda^{x_{Mj}}\right\}.$$

The closed-form MLE for λ is

$$\hat{\lambda} = (x_{.M} + x_{M.})/2n.$$

3.4 Estimation of the p and λ Parameters Under Model D

The likelihood functions for the observed data under either Model D-U or Model D-S cannot be factored and all parameter estimates must be obtained simultaneously. An iterative procedure for obtaining MLE's under Model D-U is given in Stasny (1988). The equations for this procedure are provided in Appendix II. Under Model D-S, the likelihood function for the observed data is as follows:

$$\left\{\prod_{i=1}^{K} p_{ii}^{x_{ii}}\right\} \times \left\{\prod_{i=1}^{K}\prod_{j=i+1}^{K} p_{ij}^{x_{ij}}\right\} \times \left\{\prod_{i=2}^{K}\prod_{j=i}^{i-1} p_{ji}^{x_{ij}}\right\} \times \left\{\prod_{i=1}^{K}\prod_{j=1}^{K}\left[(1-\lambda_{1i}-\lambda_{2j})\right]^{x_{ij}}\right\}$$

$$\times \left\{\prod_{i=1}^{K}\left[\sum_{j=1}^{K} p_{ij}\lambda_{2j}\right]^{x_{iM}}\right\} \times \left\{\prod_{j=1}^{K}\left[\sum_{i=1}^{K} p_{ji}\lambda_{1i}\right]^{x_{Mj}}\right\}. \qquad (2)$$

Equation (2) is maximized subject to the constraint that the sum of the p_{ij}'s is one. In general, an iterative procedure is required in order to obtain the MLE's. The iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Parameters Under Model D-S

1. $p_{ii}^{(0)} = x_{ii}/x_{..}$

$$p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..} \quad \text{for} \quad i \neq j$$

$$\lambda_{1i}^{(0)} = x_{M\cdot}/n$$

$$\lambda_{2j}^{(0)} = x_{\cdot M}/n.$$

2. $$p_{ii}^{(\nu+1)} = n^{-1}\left\{x_{ii} + x_{iM}\left[p_{ii}^{(\nu)}\lambda_{2i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{2h}^{(\nu)}\right] + x_{Mi}\left[p_{ii}^{(\nu)}\lambda_{1i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{1h}^{(\nu)}\right]\right\}$$

$$p_{ij}^{(\nu+1)} = (2n)^{-1}\left\{x_{ij} + x_{ji} + x_{iM}\left[p_{ij}^{(\nu)}\lambda_{2j}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{2h}^{(\nu)}\right]\right.$$

$$+ x_{jM}\left[p_{ij}^{(\nu)}\lambda_{2i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_{2h}^{(\nu)}\right] + x_{Mi}\left[p_{ij}^{(\nu)}\lambda_{1j}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{1h}^{(\nu)}\right]$$

$$\left. + x_{Mj}\left[p_{ij}^{(\nu)}\lambda_{1i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_{1h}^{(\nu)}\right]\right\} \quad \text{for} \quad i \neq j$$

$$\lambda_{1i}^{(\nu+1)} = \sum_{j=1}^{K}\left[x_{Mj}p_{ij}^{(\nu)}\lambda_{1i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_{1h}^{(\nu)}\right] \Big/ \sum_{j=1}^{K}\left[x_{ij}/\left(1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)}\right)\right]$$

$$\lambda_{2j}^{(\nu+1)} = \sum_{i=1}^{K}\left[x_{iM}p_{ij}^{(\nu)}\lambda_{2j}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{2h}^{(\nu)}\right] \Big/ \sum_{i=1}^{K}\left[x_{ij}/\left(1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)}\right)\right].$$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other values between zero and one satisfying the constraint that the p_{ij}'s sum to one may be used.

3.5 Estimation of the p and λ Parameters Under Model E

The likelihood functions for the observed data under either Model E-U or Model E-S cannot be factored and all parameter estimates must be obtained simultaneously. An iterative procedure for obtaining MLE's under Model E-U is given in Stasny (1988). The equations for this procedure are provided in Appendix II. Under Model E-S, the likelihood function for the observed data is as follows:

$$\left\{\prod_{i=1}^{K} p_{ii}^{x_{ii}}\right\} \times \left\{\prod_{i=1}^{K}\prod_{j=i+1}^{K} p_{ij}^{x_{ij}}\right\} \times \left\{\prod_{i=2}^{K}\prod_{j=1}^{i-1} p_{ji}^{x_{ij}}\right\} \times \left\{\prod_{i=1}^{K}\prod_{j=1}^{K} [(1-\lambda_i-\lambda_j)]^{x_{ij}}\right\}$$

$$\times \left\{\prod_{i=1}^{K}\left[\sum_{j=1}^{K} p_{ij}\lambda_j\right]^{x_{iM}}\right\} \times \left\{\prod_{j=1}^{K}\left[\sum_{i=1}^{K} p_{ji}\lambda_i\right]^{x_{Mj}}\right\}. \tag{3}$$

Equation (3) is maximized subject to the constraint that the sum of the p_{ij}'s is one. In general, an iterative procedure is required in order to obtain the MLE's. The iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Parameters Under Model E-S

1. $p_{ii}^{(0)} = x_{ii}/x_{..}$

   $p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..}$ for $i \neq j$

   $\lambda_i^{(0)} = (x_{M\cdot} + x_{\cdot M})/2n.$

2. $$p_{ii}^{(\nu+1)} = n^{-1}\left\{x_{ii} + (x_{iM} + x_{Mi})\left[p_{ii}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_h^{(\nu)}\right]\right\}$$

   $$p_{ij}^{(\nu+1)} = (2n)^{-1}\left\{x_{ij} + x_{ji} + (x_{iM} + x_{Mi})\left[p_{ij}^{(\nu)}\lambda_j^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_h^{(\nu)}\right]\right.$$

   $$\left. + (x_{jM} + x_{Mj})\left[p_{ij}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_h^{(\nu)}\right]\right\} \quad \text{for } i \neq j$$

   $$\lambda_i^{(\nu+1)} = \sum_{j=1}^{K}\left[(x_{jM} + x_{Mj})p_{ji}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_h^{(\nu)}\right]$$

   $$\Big/ \sum_{j=1}^{K}\left[(x_{ij} + x_{ji})/(1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)})\right].$$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other values between zero and one satisfying the constraint that the p_{ij}'s sum to one may be used.

4. FITS OF THE MODELS TO NCS DATA

The models described in Section 3 were fit to the NCS data described in Section 2. Recall that the NCS data for each of the years from 1975 to 1978 is summarized both by number of crimes reported in each of the two interviews during the year and by the type of crime reported. Since three survey classifications are used, we have $K = 3$. Standard errors of the parameter estimates were obtained using the observed information matrix.

Table 3a

Estimates of p_{ij} for Flows Among Number-of-Crime Classifications Under Models R, A, B, and C

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Single Crime	Multiple Crime	Crime Free	Single Crime	Multiple Crime
1975							
First Interview	Crime Free	.666 (.0075)	.098 (.0050)	.029 (.0031)	.666 (.0075)	.102 (.0035)	.032 (.0022)
	Single Crime	.106 (.0051)	.029 (.0031)	.014 (.0023)	.102 (.0035)	.029 (.0031)	.012 (.0015)
	Multiple Crime	.036 (.0032)	.011 (.0021)	.012 (.0021)	.032 (.0022)	.012 (.0015)	.012 (.0021)
1976							
First Interview	Crime Free	.669 (.0076)	.101 (.0052)	.029 (.0033)	.669 (.0076)	.099 (.0036)	.030 (.0022)
	Single Crime	.098 (.0051)	.034 (.0034)	.014 (.0025)	.099 (.0036)	.034 (.0034)	.014 (.0017)
	Multiple Crime	.031 (.0030)	.014 (.0023)	.011 (.0022)	.030 (.0022)	.014 (.0017)	.010 (.0022)
1977							
First Interview	Crime Free	.670 (.0079)	.115 (.0058)	.032 (.0034)	.671 (.0079)	.103 (.0037)	.030 (.0023)
	Single Crime	.092 (.0051)	.026 (.0032)	.016 (.0026)	.103 (.0037)	.026 (.0032)	.016 (.0018)
	Multiple Crime	.028 (.0030)	.016 (.0026)	.006 (.0017)	.030 (.0023)	.016 (.0018)	.006 (.0017)
1978							
First Interview	Crime Free	.671 (.0087)	.097 (.0062)	.027 (.0035)	.671 (.0087)	.105 (.0043)	.027 (.0025)
	Single Crime	.111 (.0061)	.032 (.0040)	.009 (.0022)	.105 (.0043)	.032 (.0040)	.010 (.0017)
	Multiple Crime	.027 (.0034)	.013 (.0027)	.013 (.0026)	.027 (.0025)	.010 (.0017)	.013 (.0026)

Note: Estimated standard errors are given in parentheses.

4.1 Estimates of the p-Parameters Under Models R, A, B, and C

Recall that the p-parameter estimates do not depend on the nonresponse mechanism under Models R, A, B, and C. For the iterative procedures used to estimate the p_{ij} under both the unconstrained and symmetric models, the criterion used for stopping the iteration was that the expected counts in the (i,j) cell of the flow matrix, $n\hat{p}_{ij}$, differed by no more than 0.5 from one step of the iterative procedure to the next. In all cases, convergence occurred rapidly, taking at most six steps. The estimates of the p_{ij} when HH's are classified by numbers of crimes reported are given in Table 3a for both the unconstrained and symmetric models. The estimates of the p_{ij} when HH's are classified by types of crimes reported are given in Table 4a for both the unconstrained and symmetric models.

Table 3b

Estimates of p_{ij} for Flows Among Number-of-Crime Classifications Under Models D-S

			Symmetric Model		
			Second Interview		
			Crime Free	Single Crime	Multiple Crime
1975	**First Interview**	Crime Free	.638 (.0104)	.106 (.0047)	.035 (.0029)
		Single Crime	.106 (.0047)	.033 (.0039)	.015 (.0019)
		Multiple Crime	.035 (.0029)	.015 (.0019)	.016 (.0027)
1976	**First Interview**	Crime Free	.645 (.0100)	.100 (.0045)	.034 (.0029)
		Single Crime	.100 (.0045)	.037 (.0041)	.017 (.0021)
		Multiple Crime	.034 (.0029)	.017 (.0021)	.015 (.0029)
1977	**First Interview**	Crime Free	.642 (.0109)	.106 (.0054)	.033 (.0032)
		Single Crime	.106 (.0054)	.031 (.0043)	.021 (.0023)
		Multiple Crime	.033 (.0032)	.021 (.0023)	.009 (.0025)
1978	**First Interview**	Crime Free	.636 (.0118)	.114 (.0056)	.028 (.0029)
		Single Crime	.114 (.0056)	.040 (.0051)	.013 (.0021)
		Multiple Crime	.028 (.0029)	.013 (.0021)	.015 (.0030)

Note: Estimated standard errors are given in parentheses.

Notice in both Tables 3a and 4a that the flow matrices of estimated probabilities under the unconstrained model for the p_{ij} appear to be fairly symmetric so that the model of symmetry in the flows is suggested as a reasonable model to consider. Also notice that the estimates of the p_{ij} do not appear to change much over the four years. The fits of these two models for the p_{ij} will be considered for each of the four models for nonresponse in Subsection 4.4 below.

Table 3c

Estimates of p_{ij} for Flows Among Number-of-Crime Classifications Under Models E-U and E-S

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Single Crime	Multiple Crime	Crime Free	Single Crime	Multiple Crime
1975							
First	Crime Free	.639 (.0104)	.102 (.0061)	.031 (.0037)	.639 (.0104)	.106 (.0047)	.035 (.0028)
Interview	Single Crime	.110 (.0061)	.033 (.0039)	.016 (.0026)	.106 (.0047)	.033 (.0039)	.015 (.0019)
	Multiple Crime	.039 (.0039)	.014 (.0025)	.016 (.0027)	.035 (.0028)	.015 (.0019)	.016 (.0027)
1976							
First	Crime Free	.645 (.0100)	.103 (.0063)	.032 (.0041)	.645 (.0100)	.101 (.0045)	.033 (.0029)
Interview	Single Crime	.098 (.0057)	.037 (.0041)	.017 (.0030)	.101 (.0045)	.037 (.0041)	.017 (.0021)
	Multiple Crime	.035 (.0037)	.017 (.0027)	.016 (.0029)	.033 (.0029)	.017 (.0021)	.016 (.0029)
1977							
First	Crime Free	.636 (.0112)	.124 (.0083)	.037 (.0050)	.642 (.0110)	.106 (.0055)	.033 (.0033)
Interview	Single Crime	.094 (.0060)	.031 (.0043)	.021 (.0031)	.106 (.0055)	.030 (.0043)	.020 (.0023)
	Multiple Crime	.029 (.0036)	.020 (.0031)	.008 (.0024)	.033 (.0033)	.020 (.0023)	.008 (.0025)
1978							
First	Crime Free	.639 (.0118)	.106 (.0078)	.029 (.0042)	.637 (.0118)	.112 (.0055)	.028 (.0029)
Interview	Single Crime	.117 (.0070)	.041 (.0051)	.011 (.0026)	.112 (.0055)	.041 (.0051)	.013 (.0021)
	Multiple Crime	.027 (.0037)	.016 (.0032)	.015 (.0030)	.028 (.0029)	.013 (.0021)	.015 (.0030)

Note: Estimated standard errors are given in parentheses.

Table 4a

Estimates of p_{ij} for Flows Among Type-of-Crime Classifications Under Models R, A, B, and C

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Property Crime	Contact Crime	Crime Free	Property Crime	Contact Crime
1975							
First	Crime Free	.666 (.0075)	.105 (.0053)	.022 (.0026)	.666 (.0075)	.111 (.0037)	.024 (.0018)
Interview	Property Crime	.118 (.0054)	.044 (.0038)	.010 (.0019)	.111 (.0037)	.044 (.0038)	.008 (.0013)
	Contact Crime	.025 (.0026)	.007 (.0016)	.004 (.0012)	.024 (.0018)	.008 (.0013)	.004 (.0012)
1976							
First	Crime Free	.669 (.0076)	.108 (.0055)	.023 (.0028)	.669 (.0021)	.108 (.0011)	.022 (.0010)
Interview	Property Crime	.108 (.0053)	.047 (.0040)	.010 (.0021)	.108 (.0011)	.047 (.0019)	.011 (.0009)
	Contact Crime	.021 (.0025)	.012 (.0021)	.002 (.0011)	.022 (.0010)	.011 (.0009)	.002 (.0012)
1977							
First	Crime Free	.670 (.0079)	.128 (.0061)	.019 (.0026)	.671 (.0078)	.115 (.0039)	.018 (.0018)
Interview	Property Crime	.103 (.0053)	.041 (.0039)	.008 (.0018)	.115 (.0039)	.041 (.0040)	.008 (.0014)
	Contact Crime	.016 (.0025)	.008 (.0021)	.006 (.0018)	.018 (.0018)	.008 (.0014)	.006 (.0017)
1978							
First	Crime Free	.671 (.0087)	.104 (.0064)	.019 (.0031)	.671 (.0088)	.112 (.0044)	.019 (.0021)
Interview	Property Crime	.119 (.0063)	.040 (.0044)	.010 (.0024)	.112 (.0044)	.040 (.0044)	.010 (.0017)
	Contact Crime	.019 (.0029)	.011 (.0025)	.006 (.0020)	.019 (.0021)	.010 (.0017)	.006 (.0020)

Note: Estimated standard errors are given in parentheses.

Table 4b

Estimates of p_{ij} for Flows Among Type-of-Crime Classifications Under Models D-S

			Symmetric Model		
			Second Interview		
			Crime Free	Property Crime	Contact Crime
1975					
	First	Crime Free	.635 (.0101)	.118 (.0046)	.026 (.0026)
	Interview	Property Crime	.118 (.0046)	.052 (.0046)	.011 (.0016)
		Contact Crime	.026 (.0026)	.011 (.0016)	.005 (.0016)
1976					
	First	Crime Free	.641 (.0098)	.110 (.0046)	.026 (.0028)
	Interview	Property Crime	.110 (.0046)	.052 (.0048)	.015 (.0021)
		Contact Crime	.026 (.0028)	.015 (.0021)	.004 (.0019)
1977					
	First	Crime Free	.642 (.0104)	.120 (.0052)	.019 (.0024)
	Interview	Property Crime	.120 (.0052)	.050 (.0049)	.011 (.0019)
		Contact Crime	.019 (.0024)	.011 (.0019)	.008 (.0022)
1978					
	First	Crime Free	.636 (.0117)	.121 (.0057)	.020 (.0025)
	Interview	Property Crime	.121 (.0057)	.049 (.0055)	.012 (.0021)
		Contact Crime	.020 (.0025)	.012 (.0021)	.008 (.0025)

Note: Estimated standard errors are given in parentheses.

Table 4c

Estimates of p_{ij} for Flows Among Type-of-Crime Classifications Under Models E-U and E-S

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Property Crime	Contact Crime	Crime Free	Property Crime	Contact Crime
1975							
First	Crime Free	.636 (.0100)	.111 (.0062)	.024 (.0034)	.636 (.0101)	.117 (.0046)	.026 (.0026)
Interview	Property Crime	.124 (.0063)	.053 (.0047)	.012 (.0023)	.117 (.0046)	.052 (.0047)	.011 (.0016)
	Contact Crime	.027 (.0033)	.009 (.0020)	.005 (.0016)	.026 (.0026)	.011 (.0016)	.005 (.0016)
1976							
First	Crime Free	.641 (.0098)	.110 (.0065)	.028 (.0041)	.641 (.0098)	.110 (.0046)	.026 (.0028)
Interview	Property Crime	.110 (.0059)	.051 (.0048)	.014 (.0028)	.110 (.0046)	.052 (.0048)	.015 (.0021)
	Contact Crime	.024 (.0033)	.016 (.0028)	.005 (.0019)	.026 (.0028)	.015 (.0021)	.005 (.0019)
1977							
First	Crime Free	.636 (.0108)	.138 (.0076)	.023 (.0035)	.641 (.0105)	.121 (.0051)	.019 (.0024)
Interview	Property Crime	.107 (.0060)	.050 (.0048)	.010 (.0022)	.121 (.0051)	.049 (.0048)	.011 (.0018)
	Contact Crime	.015 (.0028)	.011 (.0027)	.009 (.0023)	.019 (.0024)	.011 (.0018)	.009 (.0022)
1978							
First	Crime Free	.641 (.0117)	.111 (.0078)	.022 (.0040)	.640 (.0117)	.118 (.0056)	.021 (.0026)
Interview	Property Crime	.124 (.0071)	.048 (.0055)	.012 (.0029)	.118 (.0056)	.048 (.0054)	.013 (.0021)
	Contact Crime	.020 (.0033)	.014 (.0031)	.009 (.0025)	.021 (.0026)	.013 (.0021)	.008 (.0025)

Note: Estimated standard errors are given in parentheses.

4.2 Estimates of the λ-Parameters Under Models R, A, B, and C

Recall that the λ-parameter estimates under Models R, A, B, and C are the same regardless of whether the unconstrained or symmetric model is used for the p-parameters. For the iterative procedures used to estimate the λ-parameters under Models A and C, the convergence criterion used was that estimates of the λ-parameters differed by no more than .0005 from one step to the next. Convergence took between 41 and 4150 steps when it occurred in fewer than 10,000 steps after using the initial parameter estimates suggested in Appendix II. The factors of the likelihood for the observed data involving only the λ-parameters were, in some cases, not well behaved. This is particularly true for the likelihoods for the 1978 data under both Models A and C. In such cases, a grid search was used to locate appropriate starting points for the iterative procedures. A rough grid search was also used in all cases to verify that, when the iterative procedure converged, it appeared to have converged to a global rather than a local maximum.

The estimates of the λ-parameters under both the number-of-crimes and type-of-crime classifications for Models R, A, B, and C are given in Tables 5, 6, 7, and 8 respectively.

Notice that under Models R and B the estimates of the λ-parameters are the same for both the number-of-crimes and type-of-crime classifications because the probability of being a nonrespondent under those two models does not depend on survey classification. Under Models A and C, the λ-parameter estimates corresponding to the crime-free classification are the same, within rounding error, for both the number-of-crimes and type-of-crime classifications since crime-free HH's are the same under both classifications. Also notice that, under Models A and C, the λ-parameter estimates, the estimated probabilities of being a nonrespondent, generally increase as the number of victimizations or the seriousness of the crime increases.

Table 5

Estimates of λ Under Model R

	Number-of-Crimes or Type-of-Crime Classification of Data
	$\hat{\lambda}$
1975	.224 (.0035)
1976	.232 (.0035)
1977	.237 (.0036)
1978	.250 (.0040)

Note: Estimated standard errors are given in parentheses.

Table 6

Estimates of λ_{1j} and λ_{2i} Under Model A

	Number-of-Crimes Classification of Data						Type-of-Crime Classification of Data					
	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$
1975	.208	.272	.327	.221	.234	.275	.208	.280	.322	.220	.246	.246
	(.0062)	(.0159)	(.0261)	(.0064)	(.0147)	(.0242)	(.0062)	(.0151)	(.0321)	(.0064)	(.0139)	(.0303)
1976	.206*	.261*	.397*	.236*	.254*	.267*	.206	.278	.381	.235	.253	.285
	(.0063)	(.0152)	(.0268)	(.0066)	(.0153)	(.0248)	(.0063)	(.0146)	(.0327)	(.0066)	(.0144)	(.0319)
1977	.192	.263	.309	.258	.281	.326	.192	.275	.267	.258	.269	.417
	(.0064)	(.0152)	(.0265)	(.0070)	(.0171)	(.0285)	(.0064)	(.0144)	(.0327)	(.0069)	(.0159)	(.0369)
1978	.207*	.316*	.302*	.269*	.280*	.321*	.207*	.305*	.343*	.269*	.280*	.334*
	(.0072)	(.0182)	(.0308)	(.0079)	(.0176)	(.0300)	(.0072)	(.0174)	(.0364)	(.0079)	(.0166)	(.0362)

Note: * Indicates cases in which the likelihood function is not well behaved.
Estimated standard errors are given in parentheses.

Table 7

Estimates of λ_1 and λ_2 Under Model B

	Number-of-Crimes or Type-of-Crime Classification of Data	
	$\hat{\lambda}_1$	$\hat{\lambda}_2$
1975	.223	.226
	(.0058)	(.0058)
1976	.225	.240
	(.0059)	(.0060)
1977	.209	.264
	(.0059)	(.0064)
1978	.227	.273
	(.0067)	(.0071)

Note: Estimated standard errors are given in parentheses.

Table 8

Estimates of λ_i Under Model C

	Number-of-Crimes Classification of Data			Type-of-Crime Classification of Data		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
1975	.214	.252	.300	.214	.262	.284
	(.0039)	(.0118)	(.0199)	(.0039)	(.0109)	(.0262)
1976	.221	.257	.330	.221	.266	.333
	(.0040)	(.0116)	(.0210)	(.0040)	(.0109)	(.0289)
1977	.225	.271	.317	.225*	.273*	.339*
	(.0041)	(.0126)	(.0235)	(.0041)	(.0115)	(.0286)
1978	.237*	.297*	.312*	.237*	.292*	.339*
	(.0046)	(.0139)	(.0236)	(.0046)	(.0130)	(.0299)

Note: * Indicates cases in which the likelihood function is not well behaved.
Estimated standard errors are given in parentheses.

4.3 Parameter Estimates Under Models D and E

Models D and E are more difficult to fit than Models R, A, B, and C because all parameters under Models D and E must be estimated simultaneously. For all sets of the NCS data, the likelihood functions under Models D and E were not well behaved and grid searches over the possible values of the λ-parameters were required to locate suitable starting points for the iterative procedure. Since a grid search over the six λ-parameters under Model D was extremely time-consuming, parameter estimates were obtained under Model D-S but not under Model D-U. Estimates of the p-parameters under Model D-S are given in Table 3b for the number-of-crimes classification and in Table 4b for the type-of-crime classification. The λ-parameter estimates under Model D-S are given in Table 9 for both types of classifications. Estimates of the p-parameters under Models E-U and E-S are given in Table 3c for the number-of-crimes classification and in Table 4c for the type-of-crime classification. The λ-parameter estimates under Models E-U and E-S are given in Table 10 for both types of classifications.

Notice that under Models D and E the estimates of p_{11}, the probability of remaining in the crime-free classification, are somewhat smaller that the corresponding estimates under Models R, A, B, and C; the estimates of the remaining p-parameters under Models D and E are somewhat larger than the corresponding estimates under Models R, A, B, and C. Under both Models D and E, the λ-parameter estimates, the estimated probabilities of being a nonrespondent, generally increase as the number of victimizations or the seriousness of the crime increases. In the cases where the estimates decrease as the number of victimizations or the seriousness of the crime increases (in the 1978 data under Model D-S and in the 1978 number-of-crimes data under Model E-S), the decreases are small and within the estimated standard error of the estimates.

Table 9

Estimates of λ_{1i} and λ_{2j} Under Model D-S

	Number-of-Crimes Classification of Data						Type-of-Crime Classification of Data					
	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$
1975	.210	.246	.319	.194	.321	.387	.208	.264	.319	.192	.339	.372
	(.0085)	(.0303)	(.0368)	(.0085)	(.0282)	(.0362)	(.0084)	(.0249)	(.0523)	(.0085)	(.0235)	(.0507)
1976	.204	.276	.339	.217	.273	.444	.203	.280	.383	.215	.297	.453
	(.0083)	(.0274)	(.0344)	(.0084)	(.0291)	(.0331)	(.0083)	(.0244)	(.0443)	(.0084)	(.0255)	(.0416)
1977	.175	.307	.380	.249	.298	.374	.175	.304	.438	.248	.315	.341
	(.0086)	(.0301)	(.0403)	(.0089)	(.0326)	(.0439)	(.0086)	(.0243)	(.0424)	(.0089)	(.0259)	(.0491)
1978	.211	.278	.290	.236	.413	.384	.211	.276	.293	.236	.411	.391
	(.0094)	(.0282)	(.0433)	(.0099)	(.0261)	(.0443)	(.0094)	(.0264)	(.0563)	(.0098)	(.0246)	(.0567)

Note: Estimated standard errors are given in parentheses.

Table 10
Estimates of λ_i Under Model E

	Number-of-Crimes Classification of Data			Type-of-Crime Classification of Data		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
			Unconstrained p_{ij}			
1975	.202 (.0060)	.285 (.0235)	.348 (.0262)	.201 (.0058)	.302 (.0180)	.336 (.0418)
1976	.211 (.0057)	.275 (.0226)	.387 (.0232)	.209 (.0056)	.286 (.0193)	.419 (.0327)
1977	.210 (.0063)	.315 (.0259)	.372 (.0351)	.209 (.0061)	.318 (.0183)	.394 (.0295)
1978	.224 (.0065)	.340 (.0208)	.342 (.0296)	.225 (.0065)	.326 (.0203)	.385 (.0333)
			Symmetric p_{ij}			
1975	.202 (.0060)	.285 (.0235)	.351 (.0258)	.201 (.0059)	.301 (.0180)	.341 (.0408)
1976	.211 (.0057)	.274 (.0223)	.389 (.0229)	.209 (.0056)	.287 (.0191)	.418 (.0327)
1977	.213 (.0061)	.301 (.0267)	.376 (.0339)	.213 (.0060)	.309 (.0190)	.391 (.0302)
1978	.224 (.0065)	.343 (.0204)	.338 (.0298)	.225 (.0065)	.329 (.0199)	.379 (.0339)

Note: Estimated standard errors are given in parentheses.

4.4 Fits of the Models

Table 11 shows the X^2 and G^2 values and the associated degrees of freedom for all twelve models (including Model D-U which must fit the data exactly) and both types of survey classifications. Note that the models were fit as an illustration of the methods developed here and we have ignored the complex survey design. Although clusters are not a problem in our subsample of the NCS data, in a more complete analysis we would prefer to fit the models separately to data from different strata and then combine the strata estimates to obtain estimates for the entire population.

Clearly, neither Model R, the model of random nonresponse, nor Model B, under which the probability of nonresponse depends only on time, fits the data well for either the unconstrained or symmetric models for the p_{ij}.

Models C-U and C-S fit the 1975 data fairly well and give reasonable fits to the 1976 data. Since Model C-S fits the data reasonably well and is a more parsimonious model, we prefer it over Model C-U. Under Model C, the probability of nonresponse depends only on the victimization classification at the interview in which the HH responded, not on the time. Thus, Model C is the model of symmetry in the nonresponse probabilities for the two interview periods. When Model C is paired with the symmetric model for the p-parameters, we obtain symmetric expected cell counts for the observed flow data. Notice in the observed data shown in Appendix I, that in 1977 and 1978 there is much more nonresponse at the second interview time than at the first interview time. This difference in nonresponse rates is the reason for the lack of fit of Model C to the 1977 and 1978 data.

Table 11
Fits of the Models

	Number-of-Crimes Classification of Data				Type-of-Crime Classification of Data			
	Unconstrained p_{ij}		Symmetric p_{ij}		Unconstrained p_{ij}		Symmetric p_{ij}	
	X^2	G^2	X^2	G^2	X^2	G^2	X^2	G^2
Model R	(d.f. = 5)		(d.f. = 8)		(d.f. = 5)		(d.f. = 8)	
1975	42.7	41.2	45.9	45.6	38.2	36.9	42.0	41.5
1976	70.2	67.1	69.7	67.7	57.7	55.9	58.3	56.4
1977	74.2	75.2	83.9	85.3	85.4	84.8	94.8	95.3
1978	61.7	62.7	64.9	66.3	63.2	64.1	65.5	66.8
Model A	(d.f. = 0)		(d.f. = 3)		(d.f. = 0)		(d.f. = 3)	
1975	0.0	0.0	4.4	4.4	0.0	0.0	4.6	4.6
1976	0.0	0.0	0.6	0.6	0.0	0.0	0.5	0.5
1977	0.0	0.0	10.1	10.1	0.0	0.0	10.5	10.5
1978	0.0	0.0	3.7	3.7	0.0	0.0	2.7	2.7
Model B	(d.f. = 4)		(d.f. = 7)		(d.f. = 4)		(d.f. = 7)	
1975	42.7	41.1	45.9	45.5	38.2	36.9	42.0	41.5
1976	69.1	64.5	68.5	65.1	56.2	53.3	56.9	53.8
1977	47.1	45.4	58.7	55.5	57.0	54.9	68.4	65.4
1978	47.6	46.0	50.1	49.6	49.1	47.4	50.7	50.1
Model C	(d.f. = 3)		(d.f. = 6)		(d.f. = 3)		(d.f. = 6)	
1975	6.9	6.9	11.3	11.3	7.4	7.4	12.0	12.0
1976	21.2	21.3	21.8	21.9	15.1	15.1	15.6	15.6
1977	38.1	38.3	48.2	48.4	45.6	45.7	56.0	56.3
1978	31.1	31.1	34.7	34.8	29.9	30.0	32.6	32.7
Model D	(d.f. = 0)		(d.f. = 3)		(d.f. = 0)		(d.f. = 3)	
1975	0.0	0.0	5.0	5.0	0.0	0.0	5.6	5.6
1976	0.0	0.0	15.3	15.3	0.0	0.0	11.6	11.6
1977	0.0	0.0	11.5	11.5	0.0	0.0	18.0	18.0
1978	0.0	0.0	10.2	10.2	0.0	0.0	9.9	9.8
Model E	(d.f. = 3)		(d.f. = 6)		(d.f. = 3)		(d.f. = 6)	
1975	7.0	7.0	11.3	11.3	7.3	7.3	12.0	12.0
1976	21.0	21.1	21.8	21.9	14.8	14.9	15.6	15.6
1977	33.0	33.0	48.2	48.4	39.5	39.5	56.0	56.3
1978	32.0	32.1	34.6	34.8	30.9	31.0	32.6	32.7

Note: $\chi^2_{.99}(3) = 11.34$, $\chi^2_{.99}(4) = 13.28$, $\chi^2_{.99}(5) = 15.09$, $\chi^2_{.99}(6) = 16.81$, $\chi^2_{.99}(7) = 18.48$, and $\chi^2_{.99}(8) = 20.09$.

The fits of Models E-U and E-S are quite similar to those of Models C-U and C-S respectively. This is not surprising since the interpretations of the model are quite similar. Under Model C nonresponse depends on the survey classification when the HH responds while under Model E it depends on the survey classification when the HH does not respond. Since the fits of these two models are similar, we cannot choose between the two models using the data alone. Logically, Model E seems more realistic since we might expect nonresponse to depend on the current victimization status. Since the two models provide similar fits to the data, it may be that the victimization status at the time when the HH responds is generally a good indicator for the victimization status when the HH does not respond. If that is the case, we would prefer to use Model C since it is easier to fit than Model E.

Model A-S, under which nonresponse depends on both the time and on the victimization status when the HH responds fits the 1975, 1976, and 1978 data very well and gives a reasonable fit to the 1977 data. The fits of Model D-S are similar to those of Model A-S with the exception of the 1976 data which is fit much better by Model A-S. Again we cannot choose between Model A and D based on the data alone. (Models A-U and D-U fit the data exactly.) In general, we are quite pleased with the fits of Model A-S to both the number-of-crimes and type-of-crime data from all four years. Since Model A provides a reasonable fit to all the data, we conclude that nonresponse in the NCS does depend on victimization status.

Notice that, in most cases, the fits of the models as measured by X^2 and G^2 do not change much when the symmetric p_{ij} model is used rather than the unconstrained p_{ij} model. Since we gain 3 degrees of freedom going to the more parsimonious, symmetric model for the p_{ij}, we prefer this model to the unconstrained model for the p_{ij}. This choice of the symmetric model for the flow probabilities indicates that there is a certain amount of stability in victimizations reported in the first and second halves of the year in the NCS. This stability comes from the fact that symmetry in the underlying flow probabilities implies equality of marginal totals. Thus, the numbers of HH's having no crimes, one crime, or two or more crimes remain about the same from the first interview of a year to the second year. Similarly, the numbers of HH's having no crimes, a property crime, or a contact crime remain about the same from the first interview of a year to the second year.

5. CONCLUSIONS AND FUTURE WORK

We have seen that the model of symmetry in the matrices of flows among victimization classifications paired with a model under which nonresponse depends on both time and victimization status, provides a good fit to data summaries from the NCS. The same model fits the data when classification of HH's is by number of crimes reported or by type of crime reported.

The work described here is, of course, only an initial attempt to explore nonresponse and flows among victimization classifications in NCS data. For example, we noticed that the estimated symmetric probabilities of flows among the classifications did not appear to change much over the four-year period from 1975 to 1978 but the estimated probabilities of nonresponse did appear to change over this period. One might wish to fit a model to the NCS data which has constant flow probabilities but allows the nonresponse probabilities to change over time. If the nonresponse probabilities do actually change over time, not just from year to year but also from interview period to interview period, then it would be important to try to discover why these probabilities are changing.

In the work presented here, all missing data were treated the same. In fact, data may be missing because a HU rotated out of the sample, because a HH moved into or out of the sampled HU, because no one was at home, because the HH refused to respond, or for some other reason. It may be reasonable to assume that data missing because a HU rotated out of the sample is missing at random, but that other types of nonresponse are not missing at random. Stasny (1988) presents models that allow for different types of nonresponse which could be used with the models of symmetry in flows presented here. In addition, the models here do not allow for HH's which are missing at both interview periods. Since there are, of course, such HH's, one may wish to explore Markov-chain model such as those given in Stasny (1987) which do handle nonresponse at both times.

Most importantly, one may want to consider more natural summaries of the data than were used here. The data used here were summarized by first and second interview for the year. A more meaningful summary would be, say, by month or quarter of the year. If such summaries were used, then the complex nature of the interview schedule for the NCS would have to be considered and accounted for in the models. For example, the response status for a HH would be the same for the six-month reporting period covered at any one interview time. The development of models taking this into account is an important area for future work.

ACKNOWLEDGEMENTS

The National Crime Survey data utilized in this paper were made available by the Inter-University Consortium for Political and Social Research. The data were originally collected by the United States' Law Enforcement Assistance Administration. The longitudinal data set used here was created by the Bureau of Justice Statistics using the quarterly public-use data files. This research was supported in part by a grant from the Bureau of Justice Statistics, U.S. Department of Justice, and the Committee on Law and Justice Statistics, American Statistical Association, which permitted the author to attend two Workshops on the Design and Use of the National Crime Survey. The author takes sole responsibility for the work presented in this paper. The author wishes to thank the referees for their comments on an earlier version of this paper.

APPENDIX I

The Observed Data

			Classification by Number of Victimizations			
			Second Interview			
			Crime Free	Single Crime	Multiple Crime	Missing
1975						
		Crime Free	1963	256	67	901
	First	Single Crime	306	73	31	179
	Interview	Multiple Crime	95	26	24	83
		Missing	866	193	91	
1976						
		Crime Free	1884	257	53	951
	First	Single Crime	266	84	24	186
	Interview	Multiple Crime	82	34	18	75
		Missing	831	197	106	
1977						
		Crime Free	1742	260	66	994
	First	Single Crime	228	56	31	177
	Interview	Multiple Crime	63	31	10	76
		Missing	716	194	79	
1978						
		Crime Free	1370	157	45	831
	First	Single Crime	222	50	14	165
	Interview	Multiple Crime	50	18	19	66
		Missing	651	174	57	
			Classification by Type of Crime			
			Second Interview			
			Crime Free	Property Crime	Contact Crime	Missing
1975						
		Crime Free	1963	271	52	901
	First	Property Crime	331	107	22	217
	Interview	Contact Crime	70	17	8	45
		Missing	866	225	59	
1976						
		Crime Free	1884	266	44	951
	First	Property Crime	295	111	19	211
	Interview	Contact Crime	53	26	4	50
		Missing	831	235	68	
1977						
		Crime Free	1742	283	43	994
	First	Property Crime	262	89	18	194
	Interview	Contact Crime	29	12	9	59
		Missing	716	231	42	
1978						
		Crime Free	1370	173	29	831
	First	Property Crime	238	64	14	184
	Interview	Contact Crime	34	15	8	47
		Missing	651	184	47	

APPENDIX II: Procedures for Obtaining MLE's of the p and λ Parameters

Note that $x_{..} = \sum_{i=1}^{K}\sum_{j=1}^{K} x_{ij}$ is the total number of units responding at both times and $n = x_{..} + x_{.M} + x_{M.}$ is the total sample size. The starting values given below for the iterative procedures are merely suggested values. Other positive values summing to one may be used as initial values for the p-parameter estimates, and other values between zero and one may be used as initial values for the λ-parameter estimates.

MLE's for Unconstrained p_{ij}'s Under Models R, A, B, and C

1. $p_{ij}^{(0)} = x_{ij}/x_{..}$

2. $p_{ij}^{(\nu+1)} = \left[x_{ij} + x_{iM}p_{ij}^{(\nu)}/p_{i.}^{(\nu)} + x_{Mj}p_{ij}^{(\nu)}/p_{.j}^{(\nu)}\right]/n.$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the p_{ij} parameter estimates converge to a desired degree of accuracy.

MLE's for λ's Under Model A

1. $\lambda_{1j}^{(0)} = x_{M.}/n$ and $\lambda_{2i}^{(0)} = x_{.M}/n.$

2. a) $\lambda_{1j}^{(\nu+1)} = x_{Mj} \Big/ \sum_{i=1}^{K} \left[x_{ij}/\left(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)}\right)\right]$

   b) $\lambda_{2i}^{(\nu+1)} = x_{iM} \Big/ \sum_{j=1}^{K} \left[x_{ij}/\left(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)}\right)\right].$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the λ-parameter estimates converge to the desired degree of accuracy. If $x_{hM} > \sum_{j=1}^{K} x_{hj}$ or $x_{Mh} > \sum_{i=1}^{K} x_{ih}$ for some h, so that of all units responding in a particular survey classification at one interview time more did not respond at the other interview time than did respond, then the corresponding parameter estimates will, at some step, fall outside of the 0 to 1 range and alternate formulas must be used in place of those given above (see Chen and Fienberg 1974). If for some j $x_{Mj} > \sum_{i=1}^{K} x_{ij}$, then for that j, step 2a) given above is replaced by

$$\lambda_{1j}^{(\nu+1)} = 1 - \lambda_{2h}^{(\nu)} - \left(\lambda_{1j}^{(\nu)}/x_{Mj}\right)\left\{\sum_{i=1}^{K}\left[x_{ij}/\left(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)}\right)\right]\right\}\left(1 - \lambda_{1j}^{(\nu)} - \lambda_{2h}^{(\nu)}\right),$$

where h is chosen at each step of the iteration so that $\lambda_{2h}^{(\nu)} \geq \lambda_{2i}^{(\nu)}$ for all $i = 1, 2, \ldots K$. If for some i $x_{iM} > \sum_{j=1}^{K} x_{ij}$, then for that i, step 2b) given above is replaced by

$$\lambda_{2i}^{(\nu+1)} = 1 - \lambda_{1h}^{(\nu)} - \left(\lambda_{2i}^{(\nu)}/x_{iM}\right)\left\{\sum_{j=1}^{K}\left[x_{ij}/\left(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)}\right)\right]\right\}\left(1 - \lambda_{1h}^{(\nu)} - \lambda_{2i}^{(\nu)}\right),$$

where h is chosen at each step of the iteration so that $\lambda_{1h}^{(\nu)} \geq \lambda_{1j}^{(\nu)}$ for all $j = 1, 2, \ldots K$.

MLE's for λ's Under Model B

$$\hat{\lambda}_1 = x_{M\cdot}/n \quad \text{and} \quad \hat{\lambda}_2 = x_{\cdot M}/n.$$

MLE's for λ's Under Model C

1. $\lambda_i^{(0)} = (x_{iM} + x_{Mi})/2n.$

2. $$\lambda_i^{(\nu+1)} = (x_{iM} + x_{Mi}) \Big/ \left\{ \sum_{j=1}^{K} \left[(x_{ij} + x_{ji}) \big/ \left(1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)}\right) \right] \right\}.$$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the λ-parameter estimates converge to the desired degree of accuracy. If $x_{Mi} + x_{iM} > \sum_{j=1}^{K}(x_{ij} + x_{ji})$ for some i, then as for Model A an alternate formula must be used in place of step 2 above. In such cases, step 2 is replaced by

$$\lambda_i^{(\nu+1)} = 1 - \lambda_h^{(\nu)} - \left[\lambda_i^{(\nu)} \big/ (x_{iM} + x_{Mi})\right]$$

$$\left\{ \sum_{j=1}^{K} \left[(x_{ij} + x_{ji}) \big/ \left(1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)}\right) \right] \right\} \left(1 - \lambda_h^{(\nu)} - \lambda_i^{(\nu)}\right),$$

where h is chosen at each step of the iteration so that $\lambda_h^{(\nu)} \geq \lambda_j^{(\nu)}$ for all $j = 1, 2, \ldots K$.

MLE's for Parameters Under Model D-U

1. $p_{ij}^{(0)} = x_{ij}/x_{\cdot\cdot}, \quad \lambda_{1i}^{(0)} = x_{M\cdot}/n, \quad \text{and} \quad \lambda_{2j}^{(0)} = x_{\cdot M}/n.$

2. $$p_{ij}^{(\nu+1)} = n^{-1}\left\{ x_{ij} + x_{iM}\left[p_{ij}^{(\nu)}\lambda_{2j}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{2h}^{(\nu)} \right] + x_{Mj}\left[p_{ij}^{(\nu)}\lambda_{1i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{hj}^{(\nu)}\lambda_{1h}^{(\nu)} \right] \right\}$$

$$\lambda_{1i}^{(\nu+1)} = \sum_{j=1}^{K} \left[x_{Mj} p_{ij}^{(\nu)}\lambda_{1i}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{hj}^{(\nu)}\lambda_{1h}^{(\nu)} \right] \Big/ \sum_{j=1}^{K} \left[x_{ij} \big/ \left(1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)}\right) \right]$$

$$\lambda_{2j}^{(\nu+1)} = \sum_{i=1}^{K} \left[x_{iM} p_{ij}^{(\nu)}\lambda_{2j}^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_{2h}^{(\nu)} \right] \Big/ \sum_{i=1}^{K} \left[x_{ij} \big/ \left(1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)}\right) \right].$$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the λ-parameter estimates converge to the desired degree of accuracy.

MLE's for Parameters Under Model E-U

1. $p_{ij}^{(0)} = x_{ij}/x_{..}$ and $\lambda_i^{(0)} = (x_{M\cdot} + x_{\cdot M})/2n$.

2. $$p_{ij}^{(\nu+1)} = n^{-1}\left\{x_{ij} + x_{iM}\left[p_{ij}^{(\nu)}\lambda_j^{(\nu)} \Big/ \sum_{h=1}^{K} p_{ih}^{(\nu)}\lambda_h^{(\nu)}\right] + x_{Mj}\left[p_{ij}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{hj}^{(\nu)}\lambda_h^{(\nu)}\right]\right\}$$

$$\lambda_i^{(\nu+1)} = \left\{\sum_{j=1}^{K} x_{jM}\left[p_{ji}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{jh}^{(\nu)}\lambda_h^{(\nu)}\right] + x_{Mj}\left[p_{ij}^{(\nu)}\lambda_i^{(\nu)} \Big/ \sum_{h=1}^{K} p_{hj}^{(\nu)}\lambda_h^{(\nu)}\right]\right\}$$

$$\times \left\{\sum_{j=1}^{K} (x_{ij} + x_{ji})/\left(1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)}\right)\right\}^{-1}.$$

Step 2 is repeated for $\nu = 0, 1, 2, \ldots$ until the λ-parameter estimates converge to the desired degree of accuracy.

REFERENCES

CHEN, T., and FIENBERG, S.E. (1974). Two-Dimensional Contingency Tables With Both Completely and Partially Cross-Classified Data. *Biometrics*, 30, 629-642.

DODGE, R.W., and SKOGAN, W.G. (1987). The History, Organization and Utilization of the National Crime Survey. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.

FIENBERG, S.E. (1980). The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. *The Statistician*, 29, 313-350.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley & Sons.

MONTAGLIANI, H. (1987). NCS Design. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.

SAPHIRE, D.G. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey*. Lecture Notes in Statistics, Vol. 23. New York: Springer-Verlag.

STASNY, E.A. (1986). Estimating Gross Flows Using Panel Data With Nonresponse: An Example From the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A. (1987). Some Markov-Chain Models for Nonresponse in Estimating Gross Labor Force Flows. *Journal of Official Statistics*, 3, 359-373.

STASNY, E.A. (1988). Modeling Non-Ignorable Non-Response in Categorical Panel Data With an Example in Estimating Gross Labor-Force Flows. *Journal of Business and Economic Statistics*, 6, 207-219.

TAYLOR, B.T. (1987). Redesign of the National Crime Survey. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.

U.S. DEPARTMENT OF JUSTICE and BUREAU OF JUSTICE STATISTICS (1981). *The National Crime Survey: Working Papers, Volume I: Current and Historical Perspectives*. Washington, DC.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1990:

A. Agresti, *University of Florida*

J. Armstrong, *Statistics Canada*

W. Bell, *U.S. Bureau of the Census*

D. Bellhouse, *University of Western Ontario*

K. Bennett, *Statistics Canada*

J.M. Berthelot, *Statistics Canada*

D.A. Binder, *Statistics Canada*

K. Bollen, *University of North Carolina at Chapel Hill*

P.D. Bourke, *University College (Cork)*

R. Boyer, *Statistics Canada*

R. Carter, *Statistics Canada*

G.H. Choudhry, *Statistics Canada*

N. Cressie, *Iowa State University*

J.-C. Deville, *INSEE*

D. Drew, *Statistics Canada*

D.E. Duffy, *Bell Communications Research*

J.L. Eltinge, *Texas A & M University*

G. Émond, *Stone-Consolidated*

F.J. Fowler, Jr., *University of Massachusetts*

W.A. Fuller, *Iowa State University*

J.F. Gosselin, *Statistics Canada*

G. Gray, *Statistics Canada*

R.M. Groves, *U.S. Bureau of the Census*

J. Hox, *University of Amsterdam*

P. Hoyt, *Statistics Canada*

G. Kalton, *University of Michigan*

N. Laird, *Harvard University*

H. Lee, *Statistics Canada*

E. Martin, *U.S. Bureau of the Census*

S.M. Miller, *U.S. Bureau of Labor Statistics*

A.K. Nigam, *Lucknow University*

I. Munck, *Statistics Sweden*

S. Presser, *University of Maryland*

J.N.K. Rao, *Carleton University*

D.B. Rubin, *Harvard University*

K. Rust, *Westat*

I. Sande, *Bell Communications Research*

C.E. Särndal, *University of Montreal*

A. Satin, *Statistics Canada*

W.L. Schaible, *U.S. Bureau of Labor Statistics*

G. Shababb, *NPD/Nielsen, Inc.*

M. Schenker, *UCLA*

F.J. Scheuren, *U.S. Internal Revenue Service*

A. Singh, *Statistics Canada*

R. Sitter, *Carleton University*

C. Skinner, *Iowa State University*

K.P. Srinath, *Statistics Canada*

S. Sudman, *University of Illinois*

R. Tiller, *U.S. Bureau of Labor Statistics*

R.L. Winkler, *Duke University*

K. Wolter, *A.C. Nielsen*

C.F.J. Wu, *University of Waterloo*

Acknowledgements are also due to those who assisted during the production of the 1990 issues: S. Beauchamp and G. Meilleur (Photocomposition), G. Gaulin (Author Services), and M. Haight (Translation Services). Finally we wish to acknowledge J. Clarke, M. Kent, C. Lacroix, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.