Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 11, NUMBER 2
DECEMBER 1985

Canada

Statistics Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

December 1985

Published under the authority of
the Minister of Supply and
Services Canada

May 1986
8-3200-501

Price: Canada, $10.00, $20.00 a year
Other Countries, $11.50, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 11, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

EDITORIAL POLICY

The Survey Methodology Journal will publish 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, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer systems development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual 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, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two 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 $10.00 per copy, $20.00 per year in Canada, $11.50 per copy, $23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales and Services, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 11, Number 2, December 1985

CONTENTS

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 89–94
Statistics Canada

The Relationship between Statisticians and Statisticians[1]

MARTIN B. WILK[2]

I appreciate the honour of the invitation as after-dinner speaker at this 1985 annual meeting of the Statistical Society of Canada.

The honour is unfortunately accompanied by a responsability, to say something worthwhile. That is not an easy task. I thought I would approach that job in stages. So first I invented a title. Then I thought I would try to figure out what the title meant. And that was to be my speech. Regrettably, I am still unsure what the title means. But I won't let that deter me. Of course, as Yogi Berra said, "If you don't know where you're going you may not get there".

There are many people called statisticians who carry out a very diverse set of activities which are labelled statistics. In fact, at various times in my unplanned career, I have been various kinds of statisticians. That fact of language poses the question: What are the relationships among these various kinds of statisticians and statistics?

Specifically let me identify two types of statistical activity, namely probability statistics on the one hand and the work of statistical information development, carried out by statistical agencies, on the other hand. What do I mean by probabilistic statistics? Without any attempt to be precise, I mean to encompass the discipline commonly covered in standard texts and lectures including notions of analyses of variance, tests of goodness of fit, design of experiments, variance components, Bayesian estimation and so forth.

The results of the work of statistical agencies, like Statistics Canada and the Manitoba Bureau of Statistics and the U.S. Bureau of the Census, you read in the newspapers every day.

These two kinds of work are *perceived* as related, and I believe *are* related. You might say the relationship has both a *real* and an *imaginary* part – and I am not at all clear what aspects fall into which category.

Let us take a look at some of the manifestations of these two categories – which one might also label as *white collar statistics* and *blue collar statistics* (which terms are used purely to avoid laborious repetition of awkward phrases like "probability statisticians").

The Statistical Society of Canada seems to be predominantly an organization of white collar statisticians. A recent study indicated

66% academic membership
21% government agencies.

The Statistical Society of Canada lists 32 persons from Statistics Canada as members, out of 2,000 professionals.

---

[1] Invited address at the annual meeting of the Statistical Society of Canada, Winnipeg, Manitoba, June 1985.

[2] Martin B. Wilk, formerly Chief Statistician of Canada, Currently Senior Advisor to Privy Council and President of the Statistical Society of Canada.

Registration at this meeting likely consists mainly of white collar statisticians – interested primarily in the arena of probability statistics. Not only are there only a very few persons (8) from Statistics Canada, I must also report that there was only minimal interest of supervisors at Statistics Canada in sending persons to the meeting.

Let us look at examples of output from these two categories. The official journal of the Statistical Society of Canada is the Canadian Journal of Statistics. It is a quarterly. The official release announcement vehicle of Statistics Canada is the daily, which appeared 256 times last year.

A comparison of titles of publications is fascinating. For the Canadian Journal of Statistics, I selected at random fifteen key words from 122 which represented the articles published in 1983.

Here is a sample list of what white collar statisticians are writing and reading about:

Abundance distributions
Asymptotic properties
Central Wishart distribution
Chi-squared distribution
Critical values
Decision theory
Growth-curve analysis
Linear filter
Logistic process
Longitudinal studies
Multivariate linear model
Shift estimation
Spatial time series
Structural properties
Weighted least-squares estimator

Those topics are household words at this conference. But they are *not* the topics of blue collar statistical output – and many, perhaps most, blue collar statisticians would have no understanding of, or concern with, these topics, at all.

Some indication of the output of Statistics Canada is provided by the releases announced in the daily of April 29, 1985.

- total number of pigs in Canada (over 10 million)
- the number of tonnes of barley exported (over 150,000 during March 1985)
- the number of square metres of mineral wool shipped (over 6 million)

A further indication of Statistics Canada output is the table of major statistical indicators, which is updated each week in a publication, statistical highlights, sent to ministers and deputy ministers. These indicators include:

Gross National Product
Housing Starts
Bank Rate
Unemployment Rate
Consumer Price Index Increase
Weekly Earnings

And the measures relating to economic, business, trade, financial, social and labour sectors of Canadian Society.

Statistics Canada turns out statistical studies on topics such as divorce in Canada, health of Canadians, the status of women, current economic indicators, science and technology indicators, language characteristics of Canadians and so on.

I want to make it clear that I am *not* engaged in making an assessment of the relative value of these two types of outputs. Both types of work are socially desirable, as indicated by the fact each has supporting social constituencies. By definition, each is socially justified.

But what I *am* engaged in is trying to analyze the nature of relationship between these two types of activities, both of which are labelled *statistics* and carried out by people who are called *statisticians*.

We could of course simply write it off as a case of homonysm - that is the same word being used with two entirely different meanings. Or we should simply continue to ignore this discrepancy. But neither of those is wise or productive.

You are all familiar with the classic work on the advanced theory of statistics by Kendall and Stewart. Volume I involves 396 pages of text plus tables and index. These 396 pages deal with theoretical constructs of probability statistics and mathematical derivations of various formulae.

The introductory quotation to the book is attributed to O. Henry and reads as follows:

> "Let us sit on this log at the roadside", says I, "and forget the inhumanity and ribaldry of the poets. It is in the glorious columns of ascertained facts and legalized measures that beauty is to be found. In this very log we sit upon, Mrs. Sampson," says I, "is statistics more wonderful than any poem. The rings show it was sixty year old. At the depth of two thousand feet it would become coal in three thousand years. The deepest coal mine in the world is at Killingworth, near Newcastle. A box four feet long, three feet wide, and two feet eight inches deep will hold one ton of coal. If an artery is cut, compress it above the wound. A man's leg contains thirty bones. The tower of London was burned in 1841."
>
> "Go on, Mr. Pratt", says Mrs. Sampson. "Them ideas is so original and soothing. I think statistics are just as lovely as they can be." (The handbook of Hymen).

I think the quotation is lovely. And the book is, of course, an excellent example of scholarly clarity. But I do wonder what is the connection between the quotation and the text? Do the authors see a close connection? Is the quotation - which reflects work like that of the blue collar statistician-intended to justify, or motivate, the superstructure of probabilistic statistics which follows?

Do the authors believe that the constructs and formulae of their text on probabilistic statistics serve to guide or validate the work of blue collar statisticians - of statistical agencies? Or do they believe that the discipline of probability statistics is justified because its technology has been used to produce the output of statistical agencies?

What is *real* and what is *imaginary* in this relationship?

There is something of a conundrum in the relationships between the work of white collar statistics and blue collar statistics. The apparent outlook seems to be that:

- The information product is valid because it uses approved methodology.
- The methodology has status because it derives from a formulated theory.
- But the statistical theory involves constructs and mathematical logic, usually based on various unverifiable assumptions.!

What justifies the assumptions, the constructs and the theory?

In scientific work, more generally, a theory is justified as good by the usefulness of the products produced by technology derived from the theory.

Indeed, technology is often invented without *theory* and widely accepted because of its utility. Bronze and Damascus steel were developed because of their useful properties, and not because of a mathematically consistent theory of metallurgy.

To assess whether probabilistic statistics is good, we should ask whether it provides a technology to produce products that are useful and valuable.

Instead, statisticians tend to ask the inverse question, namely whether the work of blue collar statistics is valid according to the precepts of probability statistics.

Probability statistics has produced a wide variety of concepts and models and methodologies. These include areas such as:

Decision making under uncertainty
Subjective probability
Science of inference
Likelihood inference
Bayesian estimation
Time series analysis
Hypothesis testing
Tests of significance
Confidence estimation
Estimation of sampling errors
Classification methods
Regression analysis
Variance components
Design of experiments
Sample survey design
Unbiased estimators

and so on.

Many authors have asserted that the most fundamental concept in applied probabilistic statistics is the *objective assessment of uncertainty.*

But I must tell you that that notion – however appealing and philosophically profound – does not comport with the reality of the work and mandate of statistical agencies.

Let me try to establish by example the social importance of the work of blue collar statisticians. You can make a test of your own. Make a list of what you believe to be the issues of interest to Canadian Society. Your list will include matters of employment and unemployment, income of the elderly, status of women, economic growth, trade and balance of payments, family formation, population distribution, government deficit, etc.

On examination you will find that, for the large majority of such issues, your perceptions, your knowledge and your understandings depend quite directly on the statistical information produced by blue collar statisticians, *mainly* at Statistics Canada. A similar assessment would apply in any country in the world.

To emphasize this point further by a specific example, I would like to summarize some of the uses of the consumer price index.

The consumer price index is updated each month by Statistics Canada based on monthly observations of prices of a designated market basket of goods and services. The consumer price index is the most commonly used indicator of the rate of inflation. It is often referred to as the cost of living index. The consumer price index has a direct or indirect effect on nearly all Canadians. It, or individual components of which it is weighted average, is used in the calculations or definitions of income taxes, labour contracts, family allowance payments, old age security pensions, rental agreements, insurance coverage, spousal support payments, child support payments, payments to children of war veterans, student loan repayments, and many other contractual or regulatory arrangements.

To get back to the matter of objective error estimation – supposedly the central feature of probability statistics: Statistics Canada does not produce a statistical measure of the error of the consumer price index estimate. We do *not* publish interval estimates of consumer price index. We do not test the hypothesis of no change in consumer price index from month to month. We do not produce composite estimates which would supposedly reduce random error variance.

From time to time we are queried or criticized about this, even by people who are not statisticians or scientists. It seems that, having heard so often about the results of public opinion polls, members of the public have now begun to expect an error estimate to accompany published estimates. The phrase "19 times out o f 20" is now a part of the vocabulary of most newspaper readers. Of course, public opinion polls have been going on for a long time; George Gallup found a record of one taken back in 1824, when a pennsylvania newspaper published results of what was called a "straw vote taken without discrimination of parties". Modern communications and computer technology have resulted in a proliferation of polls. Because of their popularity, there has been an increase in public awareness of the fact that a statistician (or somebody.) can conduct a sample survey, make inferences, and put a measure of uncertainty on estimates.

An audit of Statistics Canada in 1983 by the Auditor General of Canada, touched on the subject of measuring the quality of statistics. The report recommended that Statistics Canada develop and disclose more measures of quality for its statistics. The agency's formal reply was that this "recommendation could not be fully implemented, since 'measures of quality' for many statistics - particularly those of a composite nature - are impossible to produce". It would be more realistic, said the Statistics Canada response, to supply "a full *description* of available information related to possible quality limitations, including, of course, quality measures when they are available".

Statistics Canada would publish more error estimates if we felt we could. It is not that we would mind admitting the possibility of error. As professor R. C. Bose used to say to his students "to err is human. Therefore, statisticians are human".

However, the usual error estimates depend on assumptions which vastly oversimplify the situation. For example, the labour force sample households, not independent individuals who have equal chances of being selected. Also, by design, the households themselves do not have equal chances of being sampled; the sampling ratio is approximately 1 in 125 at the national level, but can be as high as 1 in 24 for provinces with small populations. Can we assume, then, that all individuals are independent and have an equal probability of being unemployed? Data are gathered by means of an interview, and either the interviewer or the respondent may make an inadvertent or even a deliberate mistake. Can we ignore all possible sources of error except sampling error? Members of a given household are sampled for six consecutive months, with 1.6 of the households rotating into and out of the overall sample each month. Thus, in any month, different respondents have responded to the questionnaire different numbers of times. Can we assume that the six responses are independent over time? Sometimes, during the six months of sampling, families move away from, or move into, a particular dwelling being sampled. And, of course, there are the usual problems of non-response, outliers, and errors of data entry, computation, and printing, etc. Concern about how to handle deviations from the "usual" assumption of statistical theory is a major continuing preoccupation of some of Statistics Canada's blue and white collar statisticians.

So, on the one hand, probability statistics has contributed the appealing and important concept of objective estimation of error; and moreover the public has been educated to accept the concept and to expect it to be implemented.

On the other hand, there are many very influential and prominent statistical products produced by people called statisticians for which such measures *are not* provided, and cannot be provided at the present time.

Abstractly, there seem to be several options!

(a) Statistical agencies and probability statistics might agree to stop sharing the label "statistics" and abandon the notion of connectivity.

(b) Probability statistics could address its efforts to produce technology to deal with the reality of complex statistical information development.
(c) Statisticians might undertake a public reeducation campaign to cancel the beliefs that neat and objective measures of statistical uncertainty are possible.

As a practical matter, only option 2 can be considered. And it also holds greater promise of productive consequences for *all* statisticians of all varieties.

In an article in science last year, Ian Hacking, a philosopher of science, commented that "the quiet statisticians have changed our world – not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it".

It is gratifying to read such an assessment of the significance of probabilistic statistics as pioneered by Fisher, Neyman, Pearson, Wald and others.

But, in the vein of my topic tonight, I want to point out that there is another cadre of "quiet statisticians" – the blue collar statisticians of statistical agencies – who have also contributed to changing the world; but precisely in the manner inverse to Mr. Hacking's assessment.

Blue collar statisticians *do* discover new facts.

They *do* establish new concepts.

They *do* invent operational definitions and implement them for public consumption.

They *do* pioneer technical developments – in computing, electronic dissemination of information, computer graphics, classification systems, national accounting frameworks and so on.

Again I want to remind you my intention is not to make, or to imply, an assessment of comparative value. The issue is: what is *real* and what is *imaginary* in the relationship of blue collar statistics and white collar statistics?

Most of what blue collar statisticians do does not in reality derive from, or directly relate to, the constructs and theories and beliefs associated with probability statistics. And yet – the blue collar statisticians are somehow persuaded or coerced into paying lip service to a supposedly fundamental connectivity to those concepts.

At the same time, the white collar statisticians continue with a vague belief that if only more of the blue collar statisticians could achieve academic respectability then probability statistics would *really* impact importantly on statistical agencies.

The synergy which may be latent in the more effective relationship of the blue collar and white collar statisticians will not be developed without effort from both groups.

I don't have the wisdom to offer any revelationary proposals.

Better channels of communication are obviously needed. In that spirit, Statistics Canada has established a program of fellowships and internships.

Also in that spirit, Statistics Canada has established a network of advisory committees, including one on statistical methodology.

A number of probabilistic statisticians are on contract as consultants to Statistics Canada.

I expect there is much more opportunity for expanding seminar exchanges and working collaborations between Statistics Canada and Universities.

There is a need for improved intellectual tolerance in both groups. Perhaps the criteria and standards for publishing need to be modified.

Perhaps the basis for judging the acceptability of research grants by the Natural Sciences and Engineering Research Council of Canada should be changed.

Perhaps training programs could usefully be modified. Perhaps Statistics Canada should offer a prize for productive developments related to outstanding areas of need in the operation of statistical agencies. Maybe we should have a continuing list of the ten most wanted solutions as an incentive, and communication mode, to probability statistics researchers.

Maybe the Statistical Society of Canada should establish a tradition that every year the after-dinner speaker at the annual meeting should talk about "the relationship of statisticians and statisticians".

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 95–110
Statistics Canada

Stratification in the Canadian Labour Force Survey

J.D. DREW, Y. BÉLANGER and P. FOY[1]

ABSTRACT

The use of a multivariate clustering algorithm to perform stratification for the Labour Force Survey is described. The algorithm developed by Friedman and Rubin (1967) is modified to allow the formation of geographically contiguous strata and to delineate heterogeneous but compact primary sampling units (PSUs) within these strata. Studies dealing with stratification variables, stratification robustness over time, and type of stratification are described.

KEY WORDS: Multivariate clustering algorithm; Geographic stratification; Continuous survey.

1. INTRODUCTION

The Canadian Labour Force Survey is redesigned after every decennial census of population and housing. The redesign which occured following the 1981 Census included an intensive program of research on various aspects of the sample design (Singh, Drew and Choudhry 1984). This report describes the portion of the research program dealing with stratification methods.

Because the LFS is used not only to provide information on labour force characteristics but also as a general design for various other household surveys, one of the principal objectives of the redesign was to increase the flexibility of the LFS for general applications. Stratification was considered a means of improving efficiency for general applications, as well as variables of particular interest to the LFS, through the application of more rigorous procedures than those used in the old design.

It was therefore decided to consider the use of multivariate clustering algorithms and to compare them with the methods used in the old design. A non-hierarchical algorithm developed by Friedman and Rubin (1967) was selected on the bases of the results of evaluations of various algorithms by Judkins and Singh (1981) as part of the redesign of the Current Population Survey of the U.S. Bureau of the Census. A description of the basic algorithm and of the extensions which we have developed appears in section 2.

Sections 3 and 4 describe the evaluation studies and the stratification eventually adopted in the two main types of area distinguished by the LFS sample design, namely non-self-representing units (NSRUs) and self-representing units (SRUs). Section 4 also describes how the algorithm was adapted to delineate the primary sampling units (PSUs) within the NSR strata.

Section 5 concludes with a number of observations on the possibility of adapting the new system to other applications.

2. STRATIFICATION ALGORITHM

The basic algorithm used for stratification is a non-hierarchical multivariate algorithm developed by Friedman and Rubin (1967). This choice is based on the results of studies

[1] J.D. Drew and Y. Bélanger, Census and Household Survey Methods Division, P. Foy, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

performed by Judkins and Singh (1981) and Kostanich, Judkins, Singh and Schantz (1981), who assessed a number of stratification algorithms for the Current Population Survey of the U.S. Bureau of the Census.

The latter modified the objective function of the algorithm for sampling with probability proportional to size (PPS), and we have added the capacity to formulate compact, contiguous strata. A more complete description of the following appears in Foy (1984).

2.1 The objective function of the algorithm

The algorithm is designed to partition the stratification units (census enumeration areas) into strata which are as homogeneous as possible with respect to a number of variables of interest that is, by minimizing the sums of the squares within each stratum.

The expressions for the sums of squares in the case of sampling with PPS are shown below after introduction of the following notation:

L = number of strata to form

N = total number of units (enumeration areas)

N_k = number of units in group (stratum) k; $(N_1 + N_2 + \ldots + N_L = N)$,

T_{jk} = size measure of unit j in group k,

$T_{.k}$ = size measure of group k,

$T..$ = total size,

${}_iX_{jk}$ = observed value of variable i for unit j in group k,

${}_iX_{.k}$ = total observed values of variable i in group k,

${}_iX..$ = total observed values of variable i,

W_i = weighting factor of variable i (see section 2.4 for further details),

p = number of variables of interest.

Thus, the expression of the total sum of squares with PPS, of variable i is given by

$$SCT_i = \sum_{k=1}^{L} \sum_{j=1}^{N_k} \frac{T_{jk}}{T..} \left(\frac{T..}{T_{jk}} {}_iX_{jk} - {}_iX.. \right)^2.$$

This is also the variance expression of the estimate of ${}_iX..$ when a unit is selected with PPS. The total sum of squares weighted for all variables is thus

$$SCT = \sum_{i=1}^{p} W_i SCT_i.$$

The within-group and between-group sums of squares are obtained respectively by the following expressions:

$$SCW_i = \sum_{k=1}^{L} \frac{T..}{T_{.k}} \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{.k}} \left(\frac{T_{.k}}{T_{jk}} {}_iX_{jk} - {}_iX_{.k} \right)^2$$

and

$$SCB_i = \sum_{k=1}^{L} \frac{T_{.k}}{T_{..}} \left(\frac{T_{..}}{T_{.k}} {}_iX_{.k} - {}_iX_{..} \right)^2.$$

Their sums of squares weighted for all variables are given respectively by

$$SCW = \sum_{i=1}^{p} W_i SCW_i$$

and

$$SCB = \sum_{i=1}^{p} W_i SCB_i.$$

The within-group sum of squares of variable i, SCW_i, is also the variance expression of the estimate of ${}_iX_{..}$ when a stratum, and subsequently a unit of this stratum, is selected with PPS.

Once again, we have the following result:

$$SCT_i = SCW_i + SCB_i, \ (i = 1, \ldots, p)$$

and

$$SCT = SCW + SCB.$$

The objective function of the stratification program is SCW, the within-group sum of squares weighted for all variables. We define the stratification index for variable i, I_i, as:

$$I_i = 100 \times \frac{SCB_i}{SCT_i} \qquad i = 1, \ldots, p.$$

A high index value indicates a good clustering.

2.2 Identification of the Best Clustering

One way of identifying the best clustering would be to generate all the possible partitions of N units into L groups and then simply select the one which minimizes the objective function. This approach is rarely feasible because the number of possible partitions may be unmanageably large.

Friedman and Rubin (1967) suggest the following algorithm. Begin with any partition of the N units into L groups. Consider moving a single unit to a group other than the one it is in. Move the unit to the group which offers the greatest reduction in the objective function. If no move will produce a reduction, leave the unit where it is. Using the partition thus created, we process the second unit in the same way, then the third, etc. The application of this procedure to each unit becomes an iteration which the authors describe as a *hill-climbing pass*. After several hill-climbing passes, the algorithm reaches a point at which no move of a single unit will produce a reduction in the objective function. This point is described as a local minimum of the objective function because it is dependent on the starting partition. Another starting partition might have achieved an even lower value of the objective func-

tion. To move beyond the local minimum, Friedman and Rubin describe two procedures, the *forcing pass* and the *reassignment pass*. By applying their algorithm to data described in their article, they obtain the highest known value of the objective function 10 times out of 14 runs from different starting partitions. They use another objective function, which is maximized. With some less well-structured data, the highest value was reached in 3 out of 11 runs, although it is impossible to be certain that this is the optimal solution. In their opinion, the forcing pass and reassignment pass methods are useful only on occasion. They have more confidence in the results obtained through the use of a number of starting partitions. This view is supported by Judkins and Singh (1981). We therefore decided to use the technique involving a number of starting partitions.

Because the algorithm moves only one unit at a time, calculation of the objective function is simplified. Following the initial calculation of the objective function, we merely recalculate the contribution to the objective function of the two groups involved in the move of the unit in question.

2.3 Contiguity

Previous LFS sample designs have used strata composed of contiguous geographic units; that is, each unit in a given stratum had to be touching at least one other unit in the same stratum. One of the main reasons was the assumption that such strata would retain the efficiency of the sample design for a longer period of time than if they were formed of discontiguous units.

In order to assess this assumption and to adopt the best possible stratification, we considered two means of taking geography into account in the stratification. The first method is described by Dahmström and Hagnell (1978), and consists of the use of centroids as variables of interest. This method uses two geographic variables (centroids), which are transformations of longitude and latitude. It yields compact strata, that is, strata in which the distance between units is made minimal by minimizing the usual within-group sum of squares of the centroids. However, the minimization is tempered by minimization of the other variables of interest. Moreover, there is no assurance that these strata will be composed of contiguous units.

The other method, which we describe as the contiguity vectors approach, is new. It guarantees contiguous, but not necessarily compact, strata. Studies described in section 3 dealt with the use of each of these methods in isolation or in combination.

2.3.1 Contiguity Vectors

To ensure the formation of contiguous strata, we proceeded as follows. Optimization is performed as described in the preceding section but beginning, in this case, with a starting partition which is contiguous, and permitting the movement of unit j from stratum A to stratum B only if, in addition to reducing the sums of squares, the following conditions are met:

i) unit j is contiguous to a unit in stratum B

ii) the movement of unit j to stratum B will not disrupt the contiguity of stratum A.

In order to verify these two conditions, it is essential that we know the links of contiguity between the units. Consequently, each unit must be assigned a contiguity vector containing a list of the units contiguous to it.

The first condition is easy to verify. In order to ensure that unit j is contiguous to a unit in stratum B, we must simply find one unit in its contiguity vector which is in stratum B.

The second condition is more difficult to verify. The principle is that a stratum is said to be contiguous if each pair of units in that stratum can be connected by a contiguous chain of units in that stratum. Suppose we want to move unit j from stratum A to stratum B. We therefore have to find, for each pair of units in the contiguity vector of unit j within stratum A, another link from among the units of stratum A. At this stage, the problem becomes like finding a path through a maze.

An algorithm has also been designed to create random starting partitions whose strata are contiguous.

2.4 Weighting of Variables

The weighting factors are of particular importance, since they determine the contribution of each variable to the cluster analysis.

It is usually preferable to standardize the variables by making the weighting factors inversely proportional to the total sum of squares of each variable. This standardization makes it possible to obtain a comparable contribution by each variable to the cluster analysis.

If, after standardization, we want to assign one or more variables greater importance in relation to the other variables in the optimization, we can do so by specifying a weight greater than 1 (normal). For example, a variable with a weight of 2 would have double importance. As described in section 3.2, we tested a number of combinations of weights for the geographic and non-geographic variables in an effort to obtain compact strata without unduly affecting the minimization of the other variables.

3. STRATIFICATION IN NON-SELF-REPRESENTING UNITS

3.1 Old Design (Platek and Singh 1976)

For the purposes of the LFS, each of Canada's ten provinces is divided into a number of economic regions (ERs), consisting of areas having similar economic structures. The boundaries of the ERs are determined in consultation with the provinces. These ERs are used as primary strata. The next stage in stratification is the partition of each ER into self-representing units (SRUs) and non-self-representing units (NSRUs). The self-representing units are cities in which the expected sample is large enough to represent at least one interviewer assignment; the NSR part make up the rest of the ER. Different sample designs are used in the SRUs and the NSRUs, because the population in the NSRUs is much more widely dispersed, necessitating a larger number of sampling stages. For the same reasons, we are retaining the concept of the SRUs and the NSRUs in the redesign.

In the old design, the NSR portion of each ER was stratified into a maximum of 5 contiguous strata with a population of between 36,000 and 75,000, based on the main characteristics of the 1971 census population, as described below and as discussed at greater length by Platek and Singh (1976).

The labour force was divided into 7 categories by industry. In each ER, the three largest industries were selected on the basis of specific criteria. The unit chosen for stratification was the combined municipality, which is the geographic region enclosed within a rural municipality and as such, often contains within its boundaries urban municipalities which are geographically smaller. By comparing, for each of these units, the proportions of the labour force working in each of the three categories with the corresponding proportions at the ER level, we identified the units showing a certain similarity which were grouped into strata. This comparison was done visually with graphics. Adjustments were occasionally necessary to satisfy the size and contiguity constraints.

Within each stratum, 12 to 15 PSUs were formed, all of them representative of the stratum in terms of the stratification variables, and of the ratio of rural to urban population. The rural parts of the PSUs were formed of contiguous EAs, and the urban parts were chosen to be as near to the rural part as possible. The sizes of the strata and the PSUs were determined so that, with two PSUs per stratum, the expected sample was equivalent to one interviewer's assignment size. On the basis of these criteria, and depending on the province, the population of the PSUs varied between 3,000 and 5,000 persons. Within the PSUs, sampling occurred in 2 or 3 stages.

3.2 Studies on Stratification during Redesign

Our studies were designed to produce conclusions which would assist in certain decisions relating to the following aspects of stratification: variables to be used, types of strata (wholly rural, wholly urban, or mixed), and the importance to be assigned to contiguity. Given the very limited time available for studies prior to the formation of the new strata and PSUs, and the general expectation that contiguous strata would be preferable over time to discontiguous strata, the first two aspects were given priority.

Some experimenting was required to find the best means of achieving contiguity, either by contiguity vectors, centroids or a combination of the two. However, following the redesign, a more detailed study was undertaken on the relative desirability of contiguous versus discontiguous strata.

3.2.1 Study on Variables and Type of Stratification

One constraint on the stratification method used in the old sample design was the limited number of stratification variables which could be taken into consideration (3 per ER).

With the new algorithm, this constraint is eliminated. In addition to the seven industry variables, we wished to determine the effect caused by the use of variables relating to the survey topic, such as employment, unemployment and income, and by such characteristics as education, housing and population. The latter characteristics have proven extremely efficient in similar studies performed by the U.S. Bureau of the Census for the Current Population Survey.

Table 1 describes the various options studied with respect to the choice of variables.

As regards the type of stratification, it was decided to study the effect of having separate strata for the rural and urban parts of the ERs, as an alternative to the mixed method of the old design.

The constraints on the sample design requiring PSUs to be approximately equivalent in population size, and the ratio between rural and urban population to remain generally the same for each PSU, frequently resulted in a lack of contiguity between the rural and urban parts of the PSUs. This led to an erosion in the presumed correspondence between the PSU and the interviewer assignment. Stratification into separate rural and urban parts, which could be substratified on an optimal basis, was, it was felt, a possible solution to this problem.

The study dealt with 11 economic regions from across Canada. The strata were defined on the basis of 1971 Census data, and assessed on the basis of 1981 census data. In performing the stratification, we used the 1971 Census enumeration areas as our stratification unit, except in Quebec and Ontario. For these two provinces, we selected census subdivisions, since the large number of EAs in certain ERs (up to 400) would have made execution of the computer programs extremely costly.

We used a conversion file between the geographic units of the two censuses to perform the evaluation based on the 1981 Census. The indices based on the 1981 data were considered more appropriate for evaluation purposes, since in fact the stratification data will be an average of 7 or 8 years old for the life of the sample design. Table 2 shows the indices based on both 1971 and 1981 census data.

Table 1
Stratification Options by Variables

Variables	Stratification option				
	1	2	3	4	5
Industries (7)[a]	x	x	x	x	x
Income		x	x	x	x
Employed		x	x		x
Unemployed		x	x[b]		x
Demography (2)[c]				x	x
Housing (4)[d]				x	x
Education (1)[e]				x	x

[a] number of persons employed in agriculture, forestry and fisheries, mines manufacturing, construction, transportation, services.
[b] double weighting on unemployment.
[c] population 15-24, population 55 and over.
[d] 1-person households, 2-person households, owned dwellings, total gross rent.
[e] secondary education.

For this study, we chose to form contiguous, compact strata, using contiguity vectors and centroids with an average weight of three (see subsection 3.2.2). The number of strata per ER was the same for all options.

The following conclusions were drawn from the results of the study, which are summarized in table 2.

Type of Stratification: Rural/urban stratification was far superior to total stratification in the case of the *agriculture* variable, which is not surprising. The same phenomenon was evident for the *manufacturing* variable, although it was less spectacular. For the *income* variable, rural/urban stratification was also initially more satisfactory, but it was not particularly robust (that is, the index deteriorated over time). Rural/urban stratification was preferable for the *unemployed* variable, while there was little difference for *employed*.

Stratification Variables: Option 4, in combination with rural/urban stratification, was clearly superior for the *unemployed* variable. As regards the other variables, option 5 was slightly more satisfactory than the rest for *employed* and *income*.

3.2.2 Study on Contiguity

As previously mentioned, it was decided to retain the concept of contiguous strata for the LFS. Such strata should be better for the production of small area estimates, because of their better geographic representation. In addition, it was felt that contiguous strata would maintain the efficiency of the sample design for a long period of time.

Table 2
Stratification indices for Option

Stratification variables	Total		Rural/Urban	
	1971	1981	1971	1981
	Unemployed			
7 industries	5.4	0.1	9.9	3.8
7 industries + income + employed + unemployed	5.2	2.3	10.2	3.4
7 industries + income + employed + unemployed × 2	7.4	2.3	10.2	5.3
17 variables	6.3	6.4	11.3	4.7
15 variables (excluding employed + unemployed)	3.6	0.1	9.8	9.0
	Employed			
7 industries	2.9	0.5	8.9	4.8
7 industries + income + employed + unemployed	8.8	2.7	8.6	3.2
7 industries + income + employed + unemployed × 2	9.1	2.8	13.1	2.2
17 variables	14.1	7.8	12.2	6.4
15 variables (excluding employed + unemployed)	6.3	1.6	11.4	3.7
	Income			
7 industries	7.4	5.7	18.9	9.5
7 industries + income + employed + unemployed	11.2	6.8	22.1	5.9
7 industries + income+ employed + unemployed × 2	10.3	6.8	28.3	9.5
17 variables	10.5	9.4	24.4	11.9
15 variables (excluding employed + unemployed)	21.0	5.3	28.9	4.5
	Agriculture			
7 industries	7.4	9.7	37.0	26.0
7 industries + income + employed + unemployed	7.6	7.8	40.0	28.7
7 industries + income + employed + unemployed × 2	8.6	7.9	43.2	31.0
17 variables	6.1	1.1	40.3	31.8
15 variables (excluding employed + unemployed)	7.0	0.4	42.7	29.0
	Manufacturing			
7 industries	14.7	8.5	16.9	13.2
7 industries + income + employed + unemployed	10.9	6.6	16.5	12.1
7 industries + income + employed + unemployed × 2	5.5	4.3	14.8	16.1
17 variables	12.5	13.5	13.3	10.7
15 variables (excluding employed + unemployed)	7.2	1.4	14.1	16.4

The next question was how to use the centroids or contiguity vectors, or a combination of the two, to obtain compact, contiguous strata without allowing the geographic constraints to affect minimization of the other variables unduly.

The study was performed with the same 11 economic regions. As anticipated, the use of contiguity vectors alone resulted in strata which were contiguous, but often irregular in shape. At the same time, the use of centroids alone, even with high weights, failed to provide any guarantee of absolute contiguity.

By varying the weight of the centroids relative to the other variables, we found that a combination of a centroid weight of 3 and contiguity vectors offered a good compromise between compactness and non-geographic optimization.

3.3 Design Stratification

In view of these results and the superior results shown by a sample design using rural/urban stratification in a study on cost variance optimization (Choudhry, Lee, Drew 1985), we decided to use separate stratification for all economic regions except those in which either the rural or the urban population was too small to form at least one stratum. It was determined that each stratum should provide a sample of at least 90 dwellings, corresponding to the selection of two PSUs with a minimum take of 45 dwellings each. In cases where this requirement could not be met, we decided to proceed with overall stratification and thus to form mixed strata. This criterion led to the adoption of separate strata in over 2/3 of the ERs.

As regards the stratification variables, we compromised on a stratification based on the 15 variables of option 4 plus *employed*. *Employed* was added because its inclusion in option 4, as compared to option 5, improved the performance of the *employed* and *income* characteristics. For the same reason, *unemployed* was excluded as a stratification variable.

For the geographic constraints, it was decided to use the contiguity vectors in combination with a uniform centroid weighting of 3 in all economic regions.

A decision was also required as to the number of strata per ER. In practice, in most of the cases, there was no choice. According to the sample design, each PSU corresponds to one interviewer assignment, and we wanted to select at least two PSUs per stratum on order to produce unbiased variance estimates. Given these constraints, in almost 2/3 of the cases, only one stratum was formed with 2 or 3 selections, in the urban or rural parts or a combination of the two. In the other cases, stratification was performed in such a way as to permit the selection, again, of 2 or 3 PSUs per stratum. This decision was based on another study showing slight reductions in variance with this approach, as compared to the old sample design in which 4 to 6 PSUs were selected from each stratum (Choudhry, Lee, Drew 1985).

3.4 Study on Robustness of Contiguous and Discontiguous Strata

Robust strata are strata that maintain the efficiency of the sample design over time. Following redesign, a study was performed to determine whether contiguous strata would be more robust over time, as had been hypothesized.

The study dealt with three economic regions in Ontario, ERs 520, 540 and 580 (1981 numbering). For each of these regions, the results of the new stratification (selected for the redesign of the LFS), which consists of contiguous strata, were compared with a stratification without contiguity constraints. The strata were defined on the basis of the 1981 data, and evaluated on the basis of the 1971 data. For the contiguous strata, we used contiguity vectors with centroids, while for the discontiguous strata, we tested two options using centroid weights of 0 and 3 respectively. The stratification variables used were the same 16 variables described above (modified option 4).

The results are shown in table 3. We see that in general, the total index calculated on stratification is higher for the two options in which contiguity is not necessary, as might be expected (1981 column). However, these two options also give higher indices over time (1971 column).

Do we really need contiguous strata? Before answering this question, we would have to perform a more in depth study involving ERs from a number of provinces. Evaluation of stratification robustness would however pose certain problems. It is easy to evaluate robustness in Ontario, since stratification there is performed at the census subdivision level, which has changed very little since 1971. When stratification is performed at the level of the enumeration areas, which are very changeable, it is extremely difficult to obtain precise figures on robustness when the strata are neither compact nor contiguous.

However, should it prove that stratification without contiguity is more satisfactory, this could compensate for the possible problems involved in production of small area estimates. It could also open new horizons: once contiguity constraints are eliminated, why could we not begin by forming compact, but not necessarily contiguous, PSUs, and then grouping them into strata? This question also could only be answered by further and more detailed studies.

3.5 Formation of PSUs

The clustering algorithm was modified to permit PSU delineation in rural and mixed strata. In the rural strata in particular, the formation of the PSUs is conceptually very similar to stratification. The only difference relates to the fact that in stratification, we attempt to minimize the sums of squares of the geographic and non-geographic variables within each stratum, while in PSU formation, we want to minimize the sums of squares of the geographic variables (to obtain compact PSUs in order to reduce costs) and to maximize those of the non-geographic variables. The latter criterion enables us to obtain PSUs which are as heterogeneous as possible in terms of characteristics, so that they are all properly representative of the stratum during sampling.

There is, however, a conflict between the desired compactness of the PSUs and their heterogeneity, because of the tendency of adjacent units to possess similar characteristics. Because of low computer costs, we performed 3 delineations per stratum with centroid weights of 10, 15 and 20, relative to the other variables. The results of each delineation were then plotted on a graph whose axes are the centroids (see Figure 1). We then selected the best of the 3 delineations on the basis of the quality of variable optimization, as reflected by the stratification indices, and through reference to the graphs. A compactness index was also taken into consideration. In most cases, as it worked out, a centroid weight of 10 or 15 was selected.

Table 3

Stratification Indices by Geographic Constraints

Economic Region	No. of Strata	Geographic Constraints					
		Contiguity and Centroids (weight of 3)		Centroids (weight of 3)		None	
		1981	1971	1981	1971	1981	1971
520	2	32.2	28.5	30.2	30.1	34.5	27.0
540	3	21.8	14.1	24.9	17.8	35.2	26.8
580	4	22.8	18.9	41.4	33.7	43.7	38.5

Formation of the PSUs in the mixed strata led to an additional constraint. We wanted the proportion of urban population to be approximately the same in each PSU. Since we also wanted the PSUs to have approximately equal total populations, it was therefore necessary in some cases to split the large urban centres among a number of PSUs. The following solution was adopted:

1. The average number of parts of urban centres which a PSU will receive (N) is determined. This number depends on the proportion of the urban population in the stratum and on the number of urban units. In practice, it was set at 1 or 2. Certain strata without sufficient population or a sufficient number of urban units were reclassified as entirely rural strata.
2. The number of parts into which each urban centre will be divided is determined. The total number of parts must equal N times the number of PSUs and each urban centre is divided into a number of parts proportional to its population.

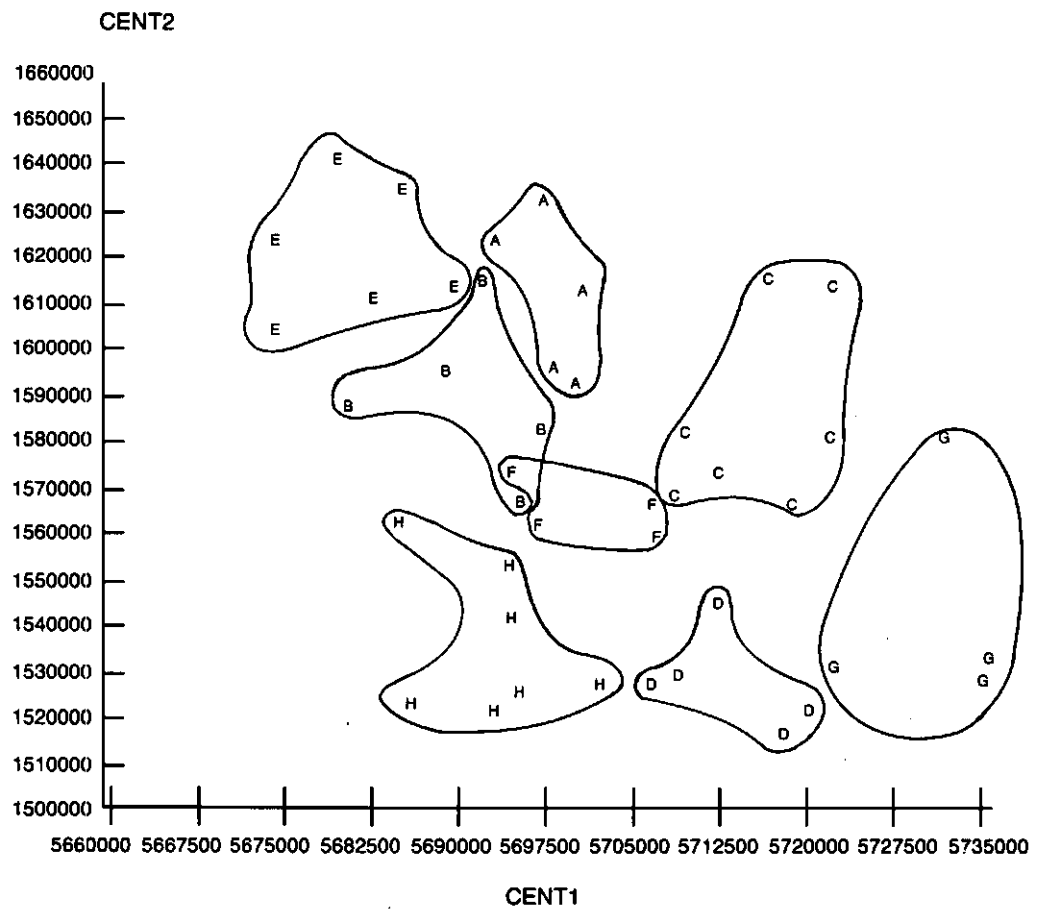


Figure 1. Example of PSU Delimitation. Each stratification unit is represented by a letter identifying the PSU to which it belongs. The PSUs are circled for clearer differentiation.

3. The optimal stratification program is applied, considering each part of an urban centre as a distinct stratification unit and adding the *urban population* variable to the other stratification variables. The weight assigned to this variable is adjusted to obtain the most evenly balanced rural/urban distribution possible within each PSU, without unduly disrupting compactness and overall optimization. This can be done only by trial and error. In practice, we found that a weight of 10 or 15 on urban population, relative to the other variables, produced satisfactory results.

In the urban strata, the PSUs were composed of urban centres. In some cases, small centres, relatively close together, were combined, without considering characteristic optimality.

Table 4 gives the average delineation indices for the PSUs in rural, mixed and urban strata. For the non-geographic variables, the lowest index represents the best delineation, while the opposite is true for the centroids. The results are clearly better, in terms of characteristic optimality, for the rural and mixed strata, in which the clustering algorithm was used. The high indices of the centroids show that the PSUs are relatively compact.

Table 4
Average PSU Delineation Indices

Variables	Type of stratum		
	Rural	Mixed	Urban
Agriculture	8.1	8.3	9.0
Forestry	21.8	24.5	35.9
Mines	20.6	36.0	57.0
Manufacturing	15.1	22.9	53.3
Construction	9.0	11.4	22.7
Transportation	9.9	12.8	22.7
Services	9.4	12.8	29.1
Employed	7.7	10.2	23.6
Unemployed[a]	13.6	14.2	18.6
Income	8.9	11.2	23.7
Population 15-24	9.4	13.4	29.8
Population 55+	7.4	13.9	34.5
1-person households	5.1	7.4	13.0
2-person households	7.9	11.9	28.1
Owned dwellings	6.8	12.5	29.4
Total gross rent	5.1	7.7	14.4
Secondary education	9.1	10.5	17.4
Total population[a]	3.2	4.0	10.5
Dwellings[a]	5.9	8.9	18.6
Centroid 1	91.6	92.7	99.2
Centroid 2	90.5	91.7	97.2

[a] Not used as a variable in optimization.

4. STRATIFICATION IN SELF-REPRESENTING UNITS

4.1 Old Design

The self-representing units of the old sample design corresponded to those cities large enough to yield an expected take equivalent to one interviewer assignment. The lower limit for SRUs varied from 10,000 persons in the Atlantic provinces to 29,000 in Quebec and Ontario.

The large SRUs were geographically stratified by grouping 3 to 5 contiguous census tracts (CTs), without any attempt to optimality. CTs are geostatistical units with populations between 3,000 and 5,000; because of their stability from one census to the next, they are practical operational units. It was felt that these strata would be efficient in estimating characteristics, and that their small size (between 10,000 and 15,000 persons) would permit sample updating in areas experiencing rapid growth, without disrupting the rest of the sample.

In addition to the area frame, an open-ended frame was set-up for apartment buildings in the large cities.

4.2 Study on stratification

Three large SRUs were considered in this study, namely Quebec City, Ottawa and Toronto. The stratification unit selected was the census tract. Because of operational constraints imposed by the stratification program, it was necessary to break Toronto up into six parts, corresponding generally to the city's major natural divisions. Stratification was carried out separately in each of these parts. The same 16 stratification variables finally selected in the NSR part were used.

Two main options were evaluated:

Option 1: Two-level stratification:
- contiguous, compact primary strata, with a centroid weighting of 3 and an expected take of approximately 150 dwellings.
- secondary strata - 4 or 5 per primary stratum, formulated without geographical constraints.

Option 2: compact stratification formulated with the use of centroids (weight of 3) and without contiguity vectors, comparable in size to the secondary strata of option 1.

Table 5 shows the results of the comparison between the old stratification and the two options studied. As in the NSR part, the strata were defined on the basis of 1971 Census data, and then evaluated on the basis of 1981 data.

We see that the two options studied consistently show better indices than the old stratification, with the possible exception of the first three variables, which, in any case, are of limited importance in cities. The old stratification nevertheless performed quite well, considering that it was carried out without any concern for optimality.

We also note that all three methods provide generally robust stratification over time, as reflected by the comparison between the indices for 1981 and 1971. Major exceptions to this rule, unfortunately, appear to be the employed and unemployed characteristics.

4.3 New Design

Given the similarity in results between the two options studied, it was decided to adopt two-level stratification (option 1) in large cities where the sample consists of 300 or more households, for the following reasons:

i) Contiguity in the primary strata gives us a suitable unit for sample updating.

ii) The primary strata can be used for the formation of interviewer assignments. The size of the strata was determined so that the sample within the geographic area, that is, the area frame sample plus the sample for the apartment frame, corresponds to two interviewer assignments (160 households in the city core and 120 elsewhere).

iii) Two-level stratification leads to better representation of the correlated response variance in variance estimates. In the old sample design, there was usually only one interviewer per stratum, resulting in an underestimate of this component of the variance. With non-geographic secondary strata, but geographic interviewer assignments, this problem will be less frequent.

The cost constraints associated with the computer time involved forced us to deal with certain SRUs on an individual basis. In fact, the Montreal region was divided into seven independent parts, during stratification. The same was done with Toronto (5 parts), Winnipeg (2 parts), Calgary (2 parts), Edmonton (2 parts) and Vancouver (3 parts). These divisions were made on the basis of natural criteria as suggested by the geography of these regions.

In large SRUs, apartment buildings existing at the time the sample design was developed were sorted by the primary strata in which they were physically located in order to achieve an implicit stratification of this sample.

Table 5

Comparison of Three Stratification Methods (SRUs)

Variables	Old Design		Two-level Stratification (Option 1)		Compact Stratification (Option 2)	
	1971	1981	1971	1981	1971	1981
Agriculture	5.5	2.9	3.2	1.8	3.4	1.8
Forestry	2.2	2.3	2.1	1.7	2.2	2.3
Mines	7.6	4.9	8.5	4.1	7.6	4.0
Manufacturing	34.7	35.0	36.6	34.1	39.1	35.0
Construction	32.5	29.6	39.7	30.1	42.4	33.4
Transportation	9.2	6.8	18.0	11.6	20.0	11.6
Services	29.5	27.5	45.8	33.1	46.7	32.1
Employed	15.1	8.0	31.4	14.1	32.8	12.6
Unemployed[a]	14.6	5.7	14.9	6.7	15.5	7.1
Income	39.4	38.6	51.8	29.8	53.6	48.0
Population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
Population 55+	27.9	18.3	34.0	20.8	32.6	18.5
1-person households	20.3	19.2	36.3	33.8	37.8	35.0
2-person households	21.9	20.3	40.3	30.9	40.1	30.2
Owned dwellings	20.3	15.3	29.7	22.9	32.1	24.9
Secondary education	32.6	42.4	50.3	47.9	51.6	49.1
Population 15+[a]	27.0	8.2	38.0	13.4	37.6	12.0
Dwellings[a]	21.8	18.5	41.7	33.8	42.1	34.3

[a]Not used as a stratification variable.

In medium-sized SRUs, where the sample was not large enough to justify two-level stratification, optimal strata were simply constructed by means of the stratification program, without the application of geographic constraints.

The smallest SRUs, those not broken into block faces for census purposes, were manually stratified, without any attempt at optimality.

Finally, we might note that the phase-in period of the new sample produced a further constraint. For large SRUs, core areas were defined as consisting of complete old-design strata that were unaffected by boundary changes. By having strata in the new design respect these core areas, we ensured that during phase-in, the new sample in core areas represented the same geographic area as the old, which permitted gradual replacement of the old sample by the new without the need for a costly parallel build up of new sample (Mayda, Drew, Lindeyer 1985).

5. CONCLUSIONS

Use of multivariate clustering algorithm enabled us to develop a very general stratification, thus strenghtening the LFS in its role as a general household survey. In addition, automation of the various stages of stratification in the NSR and SR parts, and delineation of the PSUs in the NSRUs, led to a significant reduction in the cost and time required to redesign the sample.

The system is documented (Foy 1984) and can be used for the stratification of other surveys. It may also be used in situations requiring the definition of statistical or administrative regions, using a full range of variables.

For the LFS, one aspect requiring further research relates to the selection of contiguous or discontiguous strata, and the implications of discontiguous strata on sample design.

ACKNOWLEDGEMENTS

The authors would like to thank Sylvie Trudel and Marc Joncas for their assistance in carrying out the studies mentioned in this report, and the members of the LFS Sample Redesign Committee for their valuable suggestions. They are also grateful to the referee for his helpful comments.

REFERENCES

CHOUDHRY, G.H., LEE, H., and DREW, J.D. (1985). Cost-variance optimization for the Canadian Labour Force Survey. *Survey Methodology*, 11, 33-50.

DAHMSTRÖM, P., and HAGNELL, M. (1978). The formation of strata using cluster analysis. Internal document, Department of Statistics, University of Lund, Sweden.

FOY, P. (1984). Stratification program for the Canadian Labour Force Survey: User's guide. Internal document, Census and Household Survey Methods Division, Statistics Canada.

FRIEDMAN, H.P., and RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *American Statistical Association Proceeding of the Section on Survey Research Methods*, 274-284.

KOSTANICH, D., JUDKINS, D.R., SINGH, P.R., and SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm, for use in stratified PPS sampling. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290.

MAYDA, F., DREW, J.D., and LINDEYER, J. (1985). Phase-in of the redesigned Labour Force Survey. Internal document, Census and Household Survey Methods Division, 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., and CHOUDHRY, G.H. (1984). Post '81 censal redesign of the Canadian Labour Force Survey. *Survey Methodology*, 10, 127-140.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 111–118
Statistics Canada

Sampling Microfilmed Manuscript Census Returns

D.R. BELLHOUSE[1]

ABSTRACT

In the first part of the paper a review of the historical literature concerning microfilmed manuscript census records is given. Several types of sampling designs have been used ranging in complexity from cluster and stratified random sampling to stratified two-stage cluster sampling. In the second part, a method is given to create a public use sample tape of the 1881 Census of Canada. This work was part of a pilot project for Public Archives of Canada and was carried out by the Social Science Computing Laboratory of the University of Western Ontario. The pilot project was designed to determine the merit and technical and economic feasibility of developing machine readable products from microfilm copies of the 1881 Census of Canada.

KEY WORDS: Computerized random sampling; Microfilmed records; Multi-stage designs; Public use samples; Stratification.

1. INTRODUCTION

To write a history of any person or people the historian must rely on the applicable source material. Many historians today seek to write a history of the *common man*. In this area of historical research the source material may include items such as census returns, land records, and business directories. This paper focuses on the use of census returns as a source material. The major problem with using census data is that there is a large mass of it. For an historian with a reasonable research budget there is not enough money, time or manpower to sift through all the census returns. The solution is to take a random sample of the returns. Most census returns available to the historian are microfilm copies of the returns. In Canada this includes the colonial censuses of 1841, 1851, and 1861 and the Census of Canada for 1871 and 1881. The problem then becomes one of finding the appropriate design to sample returns from the microfilm copies.

In section 2 of the paper a review of sampling techniques that have been used by historians is given. The use of sampling techniques by historians has been very uneven. Some applications have been very good; the use of a particular technique was well thought out and applied. At the other end of the spectrum other historians appear to have used overly complex designs when it was not necessary. A complex design could lead to design effects much different from 1 which, in turn, could lead to problems in the analysis of the data. See, for example, Rao and Scott (1981) and Holt *et al.* (1980) for discussions concerning categorical data analysis and Scott and Holt (1982) for regression analysis. One other problem with many of the surveys reviewed here is that there is insufficient discussion in the survey report to ascertain the reasons why a particular design was chosen.

In section 3 of the paper a method is given to sample the returns of the 1881 Census of Canada for the purpose of creating a public use sample tape. The work was carried out as part of a project for Public Archives of Canada. The contract for the research was awarded to the Social Science Computing Laboratory of the University of Western Ontario. A description of the sampling design is given here; a complete report of the project is found in Mitchell *et al.* (1982). In some ways the design is similar to the ones used for creating public

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

use sample tapes for the 1971 and 1976 Censuses of Canada. The sampling designs are all based on stratification; however, in the case of the 1881 Census, stratification could only be carried out on a geographical basis.

2. HISTORICAL REVIEW

The sampling literature for historical census documents may be categorized by the type of sampling method that was used. The order of categorization followed here will be in approximately increasing complexity of the sampling design.

2.1 Cluster Sampling

Ornstein and Darroch (1978) have given a simple cost efficient method of sampling and linking census records over time. The heart of the scheme is to form clusters of surnames and then to sample clusters. The clusters are defined by the first letter of the surname. If the same clusters are sampled over various censuses then an individual who appears in more than one census will be in the chosen sample. This reduces the number of cases to be examined for linkage purposes and hence reduces the cost. This design is particularly useful for historical studies of migration or historical changes over time.

2.2 Stratified Sampling

In all of the designs considered here that used stratification, no attempt was made to use optimal allocation. This was because prior knowledge of the variation within strata was not available to any of the researchers. To obtain the required information would have increased the cost of each project substantially.

Hammarberg (1971) used a type of two-phase or double sampling technique in an attempt to decrease the bias incurred by sampling from an incomplete set of records. The records, sampled at the second phase, were business directories for nine counties in Indiana. In the first phase of sampling, he sampled from an assumed complete record set, the 1870 United States Census. The sampling method was stratified random sampling with proportional allocation so that the sample is self-weighting. The strata were the nine counties. Two aspects of this study recur in subsequent historical sampling studies. The strata are geographical areas and the sample is self-weighting.

Hammarberg (1971) also used the classical chi-square test of fit on certain variables to see how well his sample data fit known population distributions from the census reports. In many other studies no attempt was made to check the *representativeness* of the sample.

Soltow (1975) used samples from the 1850, 1860 and 1870 United States Censuses to study wealth in the United States. For each census year he selected a sample from each microfilm reel so that the sample is stratified by reels, an approximate geographical stratification. Soltow's design appears to be a type of systematic sampling. To choose a sample he designated a spot on the screen of the microfilm reader and fed the film through the reader. The feeder arm was given successive half-turns until the manuscript census entry at the designated spot on the screen was acceptable. One criterion for sample unit selection was that the entry had to be male aged twenty years or older. Also, persons "with wealth of $100,000 or more were sampled 40 times more heavily in 1860 than those under $100,000" (p.5) so that the design is not self-weighting. Although it is not stated, the *oversampling* of wealthy people appears to have been done in order to obtain a reasonable number of them for comparison to the less affluent sections of society. Soltow (1975) also compared his sample results to the published distributions but made no statistical tests for goodness-of-fit. He found that the sample data conformed well to the census results in terms of averages and proportions on

various variables. This was true even for variables such as mean wealth, a result which is surprising in view of the oversampling of the more wealthy individuals and since his estimate appears to be the sample average.

In studying the relationship between ethnicity and occupation, Darroch and Ornstein (1980) used a sample of the 1871 Census of Canada. A description of the sampling method is given in Ornstein (1978). For the purposes of both studies it was necessary to *oversample* some ethnic groups so that the design used was not self-weighting. On ignoring the oversampling of certain ethnic groups, the sampling method used was stratified random sampling. The stratification is based on the geographical hierarchical structure of the census records : provinces, districts within provinces, sub-districts, and divisions within sub-districts. The division corresponds to the modern enumeration area. The natural stratification variable seems to be divisions. However, Ornstein (1978) further subdivided divisions into smaller groups which comprise the strata and then sampled two households per stratum. How the further subdivision was made is not given, but Ornstein states that the reason for further stratification is that sampling two units per stratum minimizes the variance of estimates of certain population values. Although it is not stated, it appears that Ornstein (1978) was trying to increase the efficiency of stratification by forming strata within a division as homogeneous as possible. By stratifying in this way the cost to sample was increased. One other aspect of Ornstein's (1978) method is that it was necessary to make at least two passes through the microfilms, the first to obtain the number of households per division and the second to sample the household.

Johnson (1978b) and Graham (1980) obtained a public use sample of the United States Census of 1900. Johnson (1978a) has described some related work in sampling the 1860 Rhode Island Census schedules. The sample was chosen by obtaining random lines on the microfilm, and then by searching for the chosen lines using a microfilm reader with an odometer attachment. Because of the sample selection procedure, the overall sample size is random. A number of criteria are given in Graham (1980, p. 41) for including or excluding sampled lines. The sampling scheme is stratified random sampling with microfilm reels as strata. The stratification is geographically based provided that the contiguous census returns are all grouped in the same microfilm. The advantages of this scheme are that it is operationally efficient and only one pass through the microfilm is needed. Also, it avoids the problem of empty strata or one unit per stratum when the sampling fraction within a stratum is small. One disadvantage is that, since one pass through the data is made, potentially major problems that arise must be dealt with on an ad hoc basis.

2.3 Stratified Cluster Sampling

Bateman and Foust (1974) obtained a sample of farms in the northern United States from the 1860 United States Census. The north was divided into two strata, East and West, and a random sample of rural counties was chosen in each stratum. Within a county one rural township (the cluster) was chosen at random and information was collected on every farm in the township. One reason for clustering appears to be due to cost considerations. The farms were obtained from the census of agriculture schedules and demographic information on the owners or operators was obtained from the census of population schedules. By remaining in the same township the work of matching farms to owners is minimized. Swierenga (1983) has provided a second reason for cluster sampling. He states that township data made it possible to estimate total factor productivity in agriculture and to identify the entire agricultural workforce, including farm laborers not residing in the 12,000 farms included in the sample (p. 793). Since the clusters, townships, were not chosen by probability proportional to size the design was not self-weighting.

Bateman and Foust (1974) also used some tests to check the representativeness of their sample. As in Hammarberg (1971), they applied the chi-square test of fit to compare sample counts to expected population counts. For continuous variables they used the t-test. The estimates of the mean and variance were the *simple* estimates, not based on the sampling design.

2.4 Stratified Two-Stage Sampling

Hammarberg (1977) used a stratified two-stage sampling scheme to sample households in the 1880 census for Utah Territory. The strata are a fairly complicated amalgamation of five geographical regions in Utah, some counties within populous regions and some large towns. Within each stratum, a sample of towns or wards was chosen. Towns which were already strata were included with certainty. Wards are geographical divisions in the Mormon Church similar to parishes in the medieval Christian church. Then a sample of households was taken from the chosen towns or wards. The sample was self-weighting on the household. The rationale for stratifying on geographical areas, given on page 460 is compelling:

> "Because the fundamental organization of the mass of people was conceived geographically, and most institutional records, – both church and secular – were organized to correspond to these areal definitions, a sample of the population on an area-by-area basis is also, in large measure, a sample of the records produced and organized for the population."

McInnis (1977) also used a stratified two-stage sampling design to obtain a sample from the 1861 Canadian Census. He studied the relationship between the number of children per family and the abundance of land in certain areas. He first stratified approximately 300 townships by their dates of settlement. Then he took a sample of townships within strata and samples of farms within townships. His reason for choosing a two-stage sample appears related to cost. A sampled farm was matched to the entry in the agricultural census. It takes less time and hence costs less to sample a few townships and match records for several farms within a township than to stratify on townships and match this record for a small number of farms. The same argument applies to Hammarberg's (1977) work. He was also linking other records to the sampled household.

2.5 Stratified Two-Stage Cluster Sampling

Smith (1978) used a stratified two-stage cluster sampling scheme to study older Americans in the 1900 United States Census. The strata are described as census regions with the counties within these regions as the primary sampling units. The primary sampling units, counties, are chosen with probability proportional to the size of their population. Within a county, several pages of census returns were sampled. Every individual over the age of 50 on each sampled page was recorded. Cluster sampling was necessary since it was too expensive to identify every individual eligible to be sampled. There is also an attempt to compare some sample distributions to the published census results. The statistic used is the standard test statistic for hypotheses on a single proportion although the data are multinomial.

A second stratified two-stage cluster sample known as the Parker-Gallman sample is described in Foust (1968, ch. 2). This sample was drawn from the 1860 Census of the United States to study the cotton growing regions in the South. The strata were 405 Southern *cotton counties*, those counties which produced 1,000 or more 400-pound bales of cotton in the year preceeding Census day. Within a county a systematic random sample of pages from the manuscript census was chosen; with a selected page a block of five farms was chosen

at random, the block being the cluster. Cluster sampling was used because information on a particular farm had to be accumulated from three different census schedules. The matching of the farms in the schedules was described as *very laborious*. Fogel and Engerman (1974, pp. 22-25) have listed several additional samples related to the Parker-Gallman sample. Bode and Ginter (1984) have criticized the content of the sample.

Of the large number of samples reviewed here, the Parker-Gallman sample and the samples drawn by Bateman and Foust (1974) are the two that have been most extensively studied. Swierenga (1983) has reviewed much of the work based on these samples.

3. PUBLIC USE SAMPLES FROM THE 1881 CENSUS OF CANADA

Early in the 1980's Public Archives of Canada obtained *Schedule 1: Nominal Return of the Living* for the 1881 Census of Canada. The returns were microfilmed and currently copies are available in most academic and many public libraries. After producing the microfilm copies, Public Archives of Canada was then interested in producing a machine readable edition of the entire census and/or a machine readable public use samples similar to the public use samples for the censuses of 1971 and 1976 (see Statistics Canada (1975, 1979) for documentation). The Social Science Computing Laboratory of the University of Western Ontario obtained a contract to perform a feasibility study and the author was asked to design a sampling scheme to construct the public use sample. In this section the proposed design is described. A report of the feasibility study is found in Mitchell *et al.* (1982).

Schedule 1 contains information on each individual on age, sex , country of birth, ethnic origin, occupation, marital status, whether or not the person had certain disabilities. The other seven schedules contain information on industry, agriculture, forestry, fishing, and mining. A brief description is found in *Census of Canada 1880-81* Vol. 1, pp. v-xv.

The basic requirements of the public use samples are briefly described. To conform to the 1971 and 1976 public use samples it would be necessary to have two independent samples, one of households and one of individuals. If production of only one sample is economically feasible, however, the first priority is the household sample. The public use sample of the 1900 Census of the United States, described by Johnson (1978b) and Graham (1980) is a sample of households. Moreover, the household appears to be the most important sampling unit desired by historians. On taking another cue from the sample of the 1900 census, a sample size in the order of one hundred thousand individuals for either the individual or the households sample is desirable. For the 1881 Census of Canada this would result in an approximate 2½% sampling fraction in either sample. Finally a stratified sampling design with proportional allocation with geographical areas as strata for both samples is desirable. This conforms to sampling practice so far in the historical literature and ensures a self-weighting design. Within a stratum the units should be chosen by simple random rather than systematic sampling. Although convenient, Johnson (1978a) has maintained that systematic sampling is not appropriate for manuscript census schedules. Neighbours possess similar characteristics and would never be included together in a systematic sample. Historians may be interested in studying those individuals with like characteristics.

Based on these basic requirements the following sampling scheme was proposed for the household sample. The design suggested was stratified random sampling with census divisions (the modern enumeration area) as strata similar to Ornstein (1978) rather than microfilm reels as used by Johnson (1978b) and Graham (1978). The census divisions provide natural geographical strata. In addition, the households are consecutively numbered on the enumerators lists with twenty-five individuals per census manuscript page. Thus, if one preliminary pass is made through the microfilms the number of households in each stratum could be easily obtained. With a 2 - 2.5% sampling fraction and proportional allocation, sample

sizes of smaller than two households are obtained in divisions (strata) with fewer than approximately one hundred households. In these cases the division should be grouped with geographically contiguous strata. Further stratification beyond the division as in Ornstein (1978) seems unnecessary and would substantially add to the sampling costs.

The sampling process can easily be made part of a computing environment. From the point of view of a coder sitting at a computer terminal with a microfilm reader to one side the sampling process is straightforward. When a coder is sampling a division, he merely presses the appropriate keys identifying the division he wants and the number of the first household to be sampled appears on the terminal screen. The coder then moves the microfilm forward to the appropriate household number. Once the data are entered, a *next* key is pressed and the second household number appears. When the final sampled household from that division is obtained, pressing the *next* key will result in an instruction to pick another division to sample. In some situations there may be missing households. For example, one or more of the enumerators sheets containing 25 names may have been lost. In this case, when a coder, in the process of sampling, encounters a missing household, the household is entered as missing and also any other missing household numbers that the coder may notice. The coder then continues sampling to the end of the division. Since at least one household sampled was missing the coder is instructed to rewind the microfilm and to continue sampling in the division. The main feature for the coder in this set-up is that with the exception of missing data situations the coder need only move the microfilm reel forward.

The computing algorithm behind this sampling method utilizes a file containing information about the divisions or division groupings and Bebbington's (1975) algorithm for drawing a simple random sample without replacement. After the initial pass is made through the microfilms a file is created containing the division identifier and the number of households in the division. If the divisions have been grouped then the size of each is recorded. When a coder identifies a division to be sampled the appropriate file entry is examined and the division size is obtained. The required sample size in the division is the division size times the sampling fraction for the whole survey which yields proportional allocation. Then Bebbington's (1975) algorithm makes a sequential choice of sample units from an ordered list, the list here being the ordered household numbers in a division. Each household number is examined in turn and is selected for or rejected from the sample. When a household number is selected the number is printed to the terminal screen and the selection procedure pauses for data entry. The sample numbers selected will be in increasing order so that a forward search only is necessary on the microfilm.

Sampling collapsed strata or grouped divisions can also be done using this algorithm. Suppose L strata of sizes $N_1, \ldots, N_L$ have been grouped into one stratum of size $N = N_1 + N_2 + \ldots + N_L$. It is necessary only to use the stratum sizes to obtain the sampled household in each stratum. Suppose in the algorithm units $s(1), \ldots, s(n)$ have been chosen for the sample, $1 \leq s(i) \leq N$. If for any $i (i = 1, \ldots, n)$ $N_1 + \ldots + N_{h-1} < s(i) \leq N_1 + \ldots + N_{h-1} + N_h$, where $N_1 + \ldots + N_{h-1} = 0$ for $h = 1$, the unit $s(i)$ is in stratum h and the household number within that stratum is $s(i) - (N_1 + \ldots + N_{h-1})$.

The general sampling algorithm can also be modified to account for missing households. The method described does not require enumerating these missing households prior to sampling. When the sampling of a stratum by Bebbington's (1975) algorithm has been completed, two possibilities arise: no missing households were encountered or some were encountered. In the former situation, there is no problem; the sampling has been completed for that situation. In the second situation, the achieved sample size, say m, is less than the desired size n. To obtain a sample of size n of the existing households, it is necessary to sample $n - m$ additional households. To achieve this, the sampling process for this stratum is started again but a list is created of the sampled and known missing households. Suppose there are M

previously sampled and known missing households ($M \geq n$: a coder may notice and record households that are missing other than those which were chosen for the sample). Define an N-dimensional vector v where the value of the u^{th} entry is u, $v(u) = u$ for $u = 1, \ldots, N$. The u^{th} entry is a pointer to the u^{th} household in a division. Now delete all entries in v corresponding to households on the microfilm which are missing or previously sampled and collapse the vector into an $(N - M)$-dimensional vector w. The values $w(u), u = 1, \ldots, N - M$ will contain the household numbers left to sample. In the algorithm, it is necessary only to restate the population size as $N - M$ and the sample size as $n - m$.

A separate and independent sample of individuals can be easily obtained using the household method of sample selection with slight modifications. The key to the modifications is that the pages of the enumerators lists are numbered with 25 names to a page. In the first pass through the microfilms, it is necessary to find the final page number and the number of lines on the last page of each division. On applying Bebbington's algorithm the computer will print the page and line number of the individual sampled.

This method of sample selection has been programmed and tested by the Social Science Computing Laboratory with positive results. For example, in the feasibility study the percentage of time spent searching for sampled units represented approximately 6% of the total estimated data entry time for the household sample and 18.5% for the individual sample. See Mitchell *et al.* (1982 pp. 20-21).

REFERENCES

BATEMAN, F., and FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.

BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.

BODE, F.A., and GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.

DARROCH, A.G., and ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.

FOGEL, R.W., and ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods.* Boston: Little, Brown and Co.

FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production.* New York: Arno Press.

GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook.* Seattle: Centre for Studies in Demography and Ecology, University of Washington.

HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.

HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880. *Journal of Interdisciplinary History*, 7, 453-476.

HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, 143, 303-320.

JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.

JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.

McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the Past*, (R.D. Lee ed.), New York: Academic Press, 201-227.

MITCHEL, S.P., LINK, D.G., and HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada.* DSS Contract Ser. No. 0SU80-00326.

ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. Unpublished manuscript, York University, Toronto.

ORNSTEIN, M.D., and DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.

RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.

SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870.* New Haven: Yale University Press.

STATISTICS CANADA (1975). *1971 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.

STATISTICS CANADA (1979). *1976 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.

SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 119–132
Statistics Canada

Estimation of Total for Two Characters in Multiple Frame Surveys

B.C. SAXENA, P. NARAIN, and A.K. SRIVASTAVA[1]

ABSTRACT

In this paper estimation of multiple characters in multiple frame surveys has been investigated. The gain due to two character study in a common survey, over separate surveys for individual characters, has been obtained. Cost comparison is also made between two character multi frame survey and two character single frame survey.

KEY WORDS: Multi-character survey; Post-stratified estimate; Optimization; Cost comparison.

1. INTRODUCTION

The technique of multiple frame surveys was suggested by Hartley (1962) and subsequently discussed by Lund (1968), Hartley (1974), Vogel (1975), Armstrong (1979), etc. Lund suggested an alternate to Hartley's estimator utilizing the actual division in the sample among various domains. Hartley (1974) further considered the problem with more general approach applicable to various sampling designs. He observed that most potential multiple frame situations employed different types of units in their respective frames. Bosecker and Ford (1976) extended Hartley's estimator to take advantage of stratification within the overlap domain. Serrurier and Phillips (1976) and Armstrong (1978) tested multiple frame techniques in agricultural surveys. The utility of multiple frame survey has been demonstrated in a wide variety of situations. In sample surveys, sometimes interest lies not only in the estimation of single character but several characters are required to be studied simultaneously. For a proper utilization of resources this is often achieved through integrated surveys. For instance, for estimating the production of vegetable crops, a single survey is planned to estimate the production of several vegetable crops. Also, besides the frame of all vegetable growers, another incomplete but relatively easily accessible frame of important vegetable growers may be utilized. In this paper, the estimation of total for two characters in multiple frame surveys has been considered. The advantage of studing more than one character in a single survey over the situation when independent surveys are planned for individual characters in a multiple frame situation, is also investigated.

2. ESTIMATOR

Let there be two overlapping frames A and B of sizes N_A and N_B respectively. In multiple frame surveys two samples of sizes n_A and n_B are selected independently by simple random sampling from frames A and B respectively. The overlapping frames generate domains a, b and ab defined as follows:

a: Consisting of units belonging to frame A only,

b: Consisting of units belonging to frame B only,

ab: Units belonging to both A and B frames.

[1] B.C. Saxena, P. Narain, and A.K. Srivastava, Indian Agricultural Statistics Research Institute, New Delhi, India.

The sample sizes n_A and n_B are split into sizes n_a, n_{ab} and n_b, n_{ba} such that n_a and n_{ab} are the number of units out of n_A units belonging to domains a and ab respectively. Similarly n_b and n_{ba} are the split of n_B units belonging to domains b and ab respectively. In the multi-character study, there will be further split of these domains generating sub-domains as follows:

Let there be two characters $y_{(1)}$ and $y_{(2)}$ under study. Then each of the usual domains a, ab and b are further subdivided as $a(1)$, $a(12)$, $a(2)$, $ab(1)$, $ab(12)$, $ab(2)$ and $b(1)$, $b(12)$, $b(2)$ respectively. Here, $a(1)$, $a(12)$ and $a(2)$ are the sub-domains consisting of units having character $y_{(1)}$, both $y_{(1)}$ and $y_{(2)}$, and $y_{(2)}$ only respectively in domain a. Similar explanation holds for other sub-domains $ab(1)$, $ab(12)$ etc. Thus the sample split in two character study will be as follows:

$$n_A = n_a + n_{ab}$$

where

$$n_a = n_{a(1)} + n_{a(2)} + n_{a(12)} \quad \text{and} \quad n_{ab} = n_{ab(1)} + n_{ab(2)} + n_{ab(12)},$$

and

$$n_B = n_b + n_{ba}$$

where

$$n_b = n_{b(1)} + n_{b(2)} + n_{b(12)} \quad \text{and} \quad n_{ba} = n_{ba(1)} + n_{ba(2)} + n_{ba(12)}.$$

Here $n_{a(1)}$, $n_{a(2)}$, etc. are the split of n_a units belonging to sub-domains $a(1)$, $a(2)$, etc. If we confine to one character then define

$$n_{A(1)} = n_{a(1)} + n_{a(12)} + n_{ab(1)} + n_{ab(12)},$$

$$n_{B(1)} = n_{b(1)} + n_{b(12)} + n_{ba(1)} + n_{ba(12)}.$$

Similarly, for the second character, $n_{A(2)}$ and $n_{B(2)}$ are defined. The estimate of the total for the first character is given by

$$\hat{Y}^{(1)} = \hat{Y}_{a(1)} + \hat{Y}^{(1)}_{a(12)} + p_1 \hat{Y}_{ab(1)} + q_1 \hat{Y}_{ba(1)} + p_2 \hat{Y}^{(1)}_{ab(12)} +$$

$$+ q_2 \hat{Y}^{(1)}_{ba(12)} + \hat{Y}_{b(1)} + \hat{Y}^{(1)}_{b(12)} \tag{1}$$

where $\hat{Y}_{a(1)}$, $\hat{Y}^{(1)}_{a(12)}$, etc . are the estimated totals for character $y_{(1)}$ of the respective sub-domains. In the subsequent discussion, for the domains in which both the characters are available, the super script corresponds to the character under consideration. For the domains having only one character the super script is not used since the domain evidently corresponds to the character.

Also, $p_1 + q_1 = 1$ and $p_2 + q_2 = 1$. Define $\bar{y}_{a(1)}$, $\bar{y}_{a(2)}$, etc. as the sample means for respective sub-domains for character $y_{(1)}$ and $y_{(2)}$ respectively.

Thus,

$$\hat{Y}^{(1)} = N_{a(1)}\bar{y}_{a(1)} + N_{a(12)}\bar{y}^{(1)}_{a(12)} + N_{ab(1)}\,(p_1\bar{y}_{ab(1)} + q_1\bar{y}_{ba(1)})$$

$$+ N_{ab(12)}(p_2\bar{y}^{(1)}_{ab(12)} + q_2\bar{y}^{(1)}_{ba(12)})$$

$$+ N_{b(12)}\bar{y}^{(1)}_{b(12)} + N_{b(1)}\bar{y}_{b(1)}. \tag{2}$$

Similarly for the second character, we have

$$\hat{Y}^{(2)} = N_{a(2)}\bar{y}_{a(2)} + N_{a(12)}\bar{y}^{(2)}_{a(12)} + N_{ab(2)}(p_3\bar{y}_{ab(2)} + q_3\bar{y}_{ba(2)})$$

$$+ N_{ab(12)}(p_4\bar{y}^{(2)}_{ab(12)} + q_4\bar{y}^{(2)}_{ba(12)}) + N_{b(12)}\bar{y}^{(2)}_{b(12)}$$

$$+ N_{b(2)}\bar{y}_{b(2)} \tag{3}$$

where

$$p_3 + q_3 = 1 \text{ and } p_4 + q_4 = 1.$$

2.1 Variance of the Estimator

The conditional variance of the post-stratified estimates $\hat{Y}^{(1)}$, $\hat{Y}^{(2)}$ for given sub-domain sample sizes ignoring the finite population correction may be written as

$$V(\hat{Y}^{(1)} \mid n_{a(1)}, n_{a(12)}, \text{ etc.}) = N^2_{a(1)}\frac{\sigma^2_{a(1)}}{n_{a(1)}} + N^2_{a(12)}\frac{\sigma^{(1)2}_{a(12)}}{n_{a(12)}}$$

$$+ N^{(2)}_{ab(1)}\left(p_1^2\frac{\sigma^2_{ab(1)}}{n_{ab(1)}} + q_1^2\frac{\sigma^2_{ab(1)}}{n_{ba(1)}}\right)$$

$$+ N^2_{ab(12)}\left(p_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ab(12)}} + q_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ba(12)}}\right) + N^2_{b(1)}\frac{\sigma^2_{b(1)}}{n_{b(1)}}$$

$$+ N^2_{b(12)}\frac{\sigma^{(1)2}_{b(12)}}{n_{b(12)}} \tag{4}$$

The unconditional variance of $\hat{Y}^{(1)}$ is approximately given by

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\left\{N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)} + p_1^2N_{ab(1)}\sigma^2_{ab(1)}\right.$$

$$\left. + p_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}\right. \tag{5}$$

$$\left. + q_1^2N_{ab(1)}\sigma^2_{ab(1)} + q_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\}$$

which is equal to the variance for stratified sampling with proportional allocation.

Similarly,

$$V(\hat{Y}^{(2)}) = \frac{N_A}{n_A}\left\{(N_{a(2)}\sigma^2_{a(2)} + N_{a(12)}\sigma^{(2)^2}_{a(12)} + p_3^2 N_{ab(2)}\sigma^2_{ab(2)}\right.$$

$$\left. + p_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(2)}\sigma^2_{b(2)} + N_{b(12)}\sigma^{(2)^2}_{b(12)}\right.$$

$$\left. + q_3^2 N_{ab(2)}\sigma^2_{ab(2)} + q_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} \tag{6}$$

where $\sigma^2_{a(1)}$, $\sigma^2_{a(2)}$, etc. are the variances for the two characters in the respective sub-domains.

For optimization of p_i's ($i = 1, 2, 3, 4$) for a common survey a combination of individual variances needs to be minimized subjet to the fixed total cost for the combined survey. Consider the simplest linear combination

$$F = V(\hat{Y}^{(1)}) + V(\hat{Y}^{(2)}).$$

For the common survey, a suitable cost function may be considered as follows:

$$C' = C_1(n_{a(1)} + n_{ab(1)}) + C_2(n_{a(12)} + n_{ab(12)}) + C_3(n_{a(2)} + n_{ab(2)})$$

$$+ C_4(n_{b(1)} + n_{ba(1)}) + C_5(n_{b(12)} + n_{ba(12)}) + C_6(n_{b(2)} + n_{ba(2)}) \tag{7}$$

where C_1 is the cost per unit in sub-domain $a(1)$, $ab(1)$; C_2 in $a(12)$, $ab(12)$; C_3 in $a(2)$, $ab(2)$ of frame A. Similarly C_4, C_5 and C_6 are the cost per unit from frame B. In the above cost function random sample sizes are involved. Consider the expected cost

$$C = E(C') = n_A(C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3) + n_B(C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6) \tag{8}$$

where

$$\Phi_1 = \frac{N_{a(1)} + N_{ab(1)}}{N_A}, \quad \Phi_2 = \frac{N_{a(12)} + N_{ab(12)}}{N_A},$$

$$\Phi_3 = \frac{N_{a(2)} + N_{ab(2)}}{N_A}, \quad \Phi_4 = \frac{N_{b(1)} + N_{ba(1)}}{N_B},$$

$$\Phi_5 = \frac{N_{b(12)} + N_{ba(12)}}{N_B}, \quad \Phi_6 = \frac{N_{b(2)} + N_{ba(2)}}{N_B}.$$

Or

$$C = n_A C_A + n_B C_B \tag{9}$$

where

$$C_A = C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3 \quad \text{and} \quad C_B = C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6.$$

In order to get the optimum p_i's as also n_A and n_B, the function F is to be minimised subject to the expected cost function as given in (9). The weight variables p_i's and sample sizes are obtained as follow using Lagrange multiplier:

$$\frac{P_1}{q_1} = \frac{P_2}{q_2} = \frac{P_3}{q_3} = \frac{P_4}{q_4} = \frac{N_B n_A}{n_B N_A} = \frac{P}{q} \text{ (say)}, \tag{10}$$

and

$$\frac{n_A^2}{N_A} = \gamma \frac{K_5 + K_1 p_1^2 + K_2 p_2^2 + K_3 p_3^2 + K_4 p_4^2}{C_A},$$

$$\frac{n_B^2}{N_B} = \gamma \frac{K_6 + K_1 q_1^2 + K_2 q_2^2 + K_3 q_3^2 + K_4 q_4^2}{C_B}, \tag{11}$$

with γ determined to meet the expected cost and

$$K_1 = N_{ab(1)}\sigma^2_{ab(1)},\ K_2 = N_{ab(12)}\sigma^{(1)2}_{ab(12)},$$

$$K_3 = N_{ab(2)}\sigma^2_{ab(2)},\ K_4 = N_{ab(12)}\sigma^{(2)2}_{ab(12)},$$

$$K_5 = N_{a(1)}\sigma^2_{a(1)} + N_{a(2)}\sigma^2_{a(2)} + N_{a(12)}(\sigma^{(1)2}_{a(12)} + \sigma^{(2)2}_{a(12)}),$$

$$K_6 = N_{b(1)}\sigma^2_{b(1)} + N_{b(2)}\sigma^2_{b(2)} + N_{b(12)}(\sigma^{(1)2}_{b(12)} + \sigma^{(2)2}_{b(12)}). \tag{12}$$

From (10) and (11), we get

$$\frac{q^2 N_B C_B}{p^2 N_A C_A} = \frac{K_6 + (K_1 + K_2 + K_3 + K_4)q^2}{K_5 + (K_1 + K_2 + K_3 + K_4)p^2} \tag{13}$$

This is a bi-quadratic in p and can be solved for p. The optimum sampling fractions can be obtained from (11). A practical case commonly met in multiple frame situations is when one of the frames has got 100% coverage. Consider 100% coverage by the frame A then $N_{b(1)} = N_{b(2)} = N_{b(12)} = 0$.

In this case (13) reduces to

$$p^2 = \frac{\alpha}{\varrho - \alpha} \frac{K_5}{K_1 + K_2 + K_3 + K_4} \tag{14}$$

where

$$\varrho = \frac{C_A}{C_B} \quad \text{and} \quad \alpha = \frac{N_B}{N_A}.$$

Assume that

$$\sigma^2_{a(1)} = \sigma^{(1)^2}_{a(12)},\ \sigma^2_{a(2)} = \sigma^{(2)^2}_{a(12)},\ \sigma^2_{ab(1)} = \sigma^{(1)^2}_{ab(12)},\ \sigma^2_{ab(2)} = \sigma^{(2)^2}_{ab(12)}. \tag{15}$$

These assumptions appear plausible since the variability of one character is not likely to be affected by the presence or absence of the other character. Then p^2 reduces to

$$p^2 = \frac{\alpha}{\varrho - \alpha}\left\{\frac{\sigma^2_{a(1)}(N_{a(1)} + N_{a(12)}) + \sigma^2_{a(2)}(N_{a(2)} + N_{a(12)})}{\sigma^2_{ab(1)}(N_{ab(1)} + N_{ab(12)}) + \sigma^2_{ab(2)}(N_{ab(2)} + N_{ab(12)})}\right\}$$

or

$$p^2 = \frac{(1 - \alpha)\Phi_2'}{(\varrho - \alpha)}\left\{\frac{\Phi_3'(\xi_1 + \xi_2) + (1 - \xi_1)}{\Phi_4'(\xi_3 + \xi_4) + (1 - \xi_3)}\right\} \tag{16}$$

where

$$\Phi_1' = \frac{\sigma^2_{a(1)}}{\sigma^2_{ab(1)}},\ \Phi_2' = \frac{\sigma^2_{a(2)}}{\sigma^2_{ab(2)}},\ \Phi_3' = \frac{\sigma^2_{a(1)}}{\sigma^2_{a(2)}},\ \Phi_4' = \frac{\sigma^2_{ab(1)}}{\sigma^2_{ab(2)}}$$

and

$$\xi_1 = \frac{N_{a(1)}}{N_a},\ \xi_2 = \frac{N_{a(12)}}{N_a},\ \xi_3 = \frac{N_{ab(1)}}{N_{ab}},\ \xi_4 = \frac{N_{ab(12)}}{N_{ab}}.$$

Using that $N_{ab} = N_B$, $N_{a(2)} + N_{a(12)} = N_a - N_{a(1)}$ and $N_{ab(2)} + N_{ab(12)} = N_{ab} - N_{ab(1)}$, it may be seen that the above expression of p^2 reduces to the usual form in uni-character case since $\xi_1 = \xi_3 = 1$ and $\xi_2 = \xi_4 = 0$. It may be remarked that the domain variances are generally not known as such these values are based either on prior knowledge or some guessed values. The optimality of p^2 is effected to that extent.

3. COMPARISON OF MULTI-CHARACTER SURVEY WITH INDEPENDENT UNI-CHARACTER SURVEYS IN MULTIPLE FRAME SITUATIONS

Multi-character surveys are planned with a view to economise the available resources and it is expected that a common survey is likely to score over independent uni-character surveys taking into account the cost and efficiency. In this situation the extent of gain due to a common multiple frame survey is investigated.

In a single character study for character $y_{(1)}$ (say), consider simple random samples of sizes n_A and n_B from the frames A and B respectively. Here we assume that the only frames used before are available, not the reduced frame for each character. Define N_{A1}, N_{B1}, n^*_{A1}, and n^*_{B1} as the population sizes and sample sizes respectively with character $y_{(1)}$. Here,

n^*_{A1} and n^*_{B1} are the random sample sizes with $E(n^*_{A1}) = n_A N_{A1}/N_A$ and $E(n^*_{B1}) = n_B N_{B1}/N_B$. In this case, the estimator $\hat{Y}^{(1)*}$ and its variance are as follows:

$$\hat{Y}^{(1)*} = (N_{a(1)} + N_{a(12)})\bar{y}_{(a(1),\, a(12))}$$

$$+ (N_{ab(1)} + N_{ab(12)})(p'\bar{y}_{(ab(1),\, ab(12))} + q'\bar{y}_{(ba(1),\, ba(12))})$$

$$+ (N_{b(1)} + N_{b(12)})\bar{y}_{(b(1),\, b(12))}$$

where p', q' are weight variables such that $p' + q' = 1$ and $\bar{y}_{(a(1),\, a(12))}, \bar{y}_{(ab(1),\, ab(12))}$, etc. are sample means for the sample from combined respective domains, e.g. $\bar{y}_{(a(1),\, a(12))}$ is the mean of sample units coming from domain $a(1)$ and $a(12)$.

$$V(\hat{Y}^{(1)*}) = \frac{N_A}{n_A}(N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)}$$

$$+ (p'^2\frac{N_A}{n_A} + q'^2\frac{N_B}{n_B})(N_{ab(1)}\sigma^2_{ab(1)} + N_{ab(12)}\sigma^{(1)2}_{ab(12)})$$

$$+ \frac{N_B}{n_B}(N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}). \tag{17}$$

In this case, the cost function is of the form

$$C = C_1 n^*_{A1} + C_4 n^*_{B1}$$

and expected cost is given by C^* as

$$C^* = C_1\frac{n_A}{N_A}N_{A1} + C_4\frac{n_B}{N_B}N_{B1} = C'_A n_A + C'_B n_B \tag{18}$$

where $C'_A = C_1 N_{A1}/N_A$ and $C'_B = C_4 N_{B1}/N_B$.

For simplicity, we assume 100% coverage by frame A, equality of variances as in (15), and $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$. Based on these assumptions, the cost C^* with n_A and n_B which minimize the variance (17) is given by (see Appendix for derivation).

$$C^* = \frac{(\xi_1 + \xi_2)[\{C_1(1 + \alpha^*_1)(\Phi'_1 + \alpha^*_1 p'^2)\}^{1/2} + \alpha^*_1(C_4 q'^2)^{1/2}]^2}{\frac{1}{1-\alpha}\left\{\frac{(\Phi'_1 + \alpha^*_1 p^2)}{n_A} + \frac{\alpha\alpha^*_1 q^2}{n_B}\right\}}$$

where

$$\alpha_1^* = \frac{\alpha}{1-\alpha}\frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

Similarly, for the separate survey for the 2nd character, the cost is obtained as

$$C^{**} = \frac{(1-\xi_1)\left[\left\{C_3(1+\alpha_2^*)(\Phi_2' + \alpha_2^* p''^2)\right\}^{1/2} + \alpha_2^*(C_6 q''^2)^{1/2}\right]^2}{\frac{1}{1-\alpha}\left\{\frac{(\Phi_2' + \alpha_2^* p^2)}{n_A} + \frac{\alpha\alpha_2^* q^2}{n_B}\right\}} \tag{19}$$

where

$$p''^2 = \frac{K\Phi_2'}{1 + \alpha_2^*(1-K)}, \quad \alpha_2^* = \frac{\alpha}{1-\alpha}\frac{1-\xi_3}{1-\xi_1}.$$

For the combined character study, the total cost C for 100% coverage by the frame A is given by (8).

Thus

$$C = \frac{n_A}{N_A}\left[C_1(N_{a(1)} + N_{ab(1)}) + C_2(N_{a(12)} + N_{ab(12)}) + C_3(N_{a(2)} + N_{ab(2)})\right]$$

$$+ \frac{n_B}{N_B}\left[C_4 N_{ab(1)} + C_5 N_{ab(12)} + C_6 N_{ab(2)}\right].$$

Using assumptions in costs (i.e. $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$) we get

$$C = C_2 n_A\left[(1-\alpha)\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} +\right.$$

$$\left.\alpha\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\} + \frac{K}{r}\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\right] \tag{20}$$

where $r = n_A/n_B$, $\varrho_1 = C_1/C_2$ *and* $\varrho_3 = C_3/C_2$.

But in combined character study (n_A/n_B) Opt. $= p/\alpha q$ where p is given by (16). Thus the gain may be obtained from the ratio.

$$\frac{C^* + C^{**}}{C} =$$

$$\frac{\frac{(\xi_1+\xi_2)\varrho_1 T_1^2}{(\Phi_1' + \alpha_1^* p)} + \frac{(1-\xi_1)\varrho_3 T_2^2}{(\Phi_3' + \alpha_2^* p)}}{\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} + \left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\left\{\frac{r\alpha + K}{r(1-\alpha)}\right\}} \tag{21}$$

where

$$T_1 = \left\{(\Phi_1' + \alpha_1^* p'^2)(1 + \alpha_1^*)\right\}^{1/2} + \alpha_1^* q' \sqrt{K}$$

$$T_2 = \left\{(\Phi_2' + \alpha_2^* p''^2)(1 + \alpha_2^*)\right\}^{1/2} + \alpha_2^* q'' \sqrt{K}.$$

K can be determined as follows: Using the definitions of C_A, C_B, Φ_i's ($i = 1, \ldots, 6$) and equation (A.1), we obtain

$$\frac{C_A}{C_B} = \frac{1}{K} \frac{\varrho_1 \Phi_1 + \Phi_2 + \varrho_3 \Phi_3}{\varrho_1 \Phi_4 + \Phi_5 + \varrho_3 \Phi_6} = \varrho,$$

and thus

$$K = \varrho^{-1}\left\{\alpha + (1 - \alpha)\frac{\varrho_1 \xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2)}{\varrho_1 \xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4)}\right\}. \tag{22}$$

The expression in (21) may be used to obtain the gain in cost due to studying both the character simultaneously in comparison to independent individual surveys. The percent gain G is thus given by

$$G = \left(\frac{C^* + C^{**}}{C} - 1\right) \times 100$$

In the above cost comparison, the expected costs, C, C^* and C^{**} do not include the overhead costs for the combined or individual surveys, however, it is expected that the sum of overhead costs pertaining to individual surveys would be much larger than the corresponding overhead cost for the combined survey. Therefore, the actual gain in costs due to common multiple frame surveys compared to independent surveys will be larger than the percent gain G defined above.

The expression (21) reduced substantially under the assumptions $\Phi_1' = \Phi_2' = \Phi$ (say) and $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi$ (say).

From (22) $\varrho = 1/K$ and from (16) since $\Phi_1'/\Phi_2' = \Phi_3'/\Phi_4'$, the p^2 reduces as follows:

$$p^2 = \frac{K(1 - \alpha)}{1 - k\alpha} \Phi.$$

Also $\alpha_1^* = \alpha_2^* = \alpha/(1 - \alpha)$.

Therefore, from (A.1)

$$p' = p'' = \left\{\frac{K(1 - \alpha)\Phi}{1 - K\alpha}\right\}^{1/2}.$$

Thus

$$T_1 = T_2 = \left\{\Phi \frac{1 - K\alpha}{1 - \alpha}\right\}^{1/2} + \frac{\alpha \sqrt{K}}{1 - \alpha}.$$

With all these substitutions in (21) $(C^* + C^{**})/C$ simplifies as follows:

$$\frac{C^* + C^{**}}{C} = \frac{T_1^2}{\Phi + \alpha_1^* p} \times \frac{2\xi\varrho_1 + (1 - \xi)\varrho_3}{\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}\{\varrho_1\xi + \xi + \varrho_3(1 - 2\xi)\}}$$

$$= \frac{T_1^2}{(\Phi + \alpha_1^* p)\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}} \times \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(\varrho_1 + 1 - 2\varrho_3)}$$

$$= \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)}$$

where $r = (n_A/n_B)$ opt. $= p/\alpha q$ from (10).

Hence,

$$G = \frac{\xi(\varrho_1 + \varrho_3 - 1)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)} \times 100.$$

The equality of ξ_i's does not seem to be realistic assumption. The value of G, has therefore been calculated using (19) for realistic and representative combinations of parameters and are presented in Table. 1.

This table indicates that there is a definite gain due to integration of multiple frame surveys for both the characters in comparison to separate individual surveys. The gain increases with increasing values of ϱ_1 and ϱ_3.

4. COMPARISON OF TWO CHARACTER MULTIPLE FRAME SURVEYS WITH SINGLE FRAME SURVEY

Comparison of two frame survey with single frame surveys for study of two characters is of practical interest. For single character a similar study was carried out by Hartley (1962). On similar lines the relative reduction in cost was obtained as

$$R = \left(1 + \frac{\alpha q}{p\varrho}\right)^2 \Big/ \left\{1 + \frac{\alpha q(1 + p)}{p^2\varrho}\right\}$$

where p^2 is given by (16), $\varrho = C_A/C_B$ and $\alpha = N_{ab}/N_A$

The reduction in cost due to multiple frame over a single frame survey is tabulated in Table 2 for some set of parametric values. The table indicates considerable cost reduction.

Table 1

Percent Gain in Cost for Common multiple Frame Survey for Both Characters over Individual Surveys,

When $\varrho = 10$, $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, $\alpha = 0.5$.

	ϱ_3						
ϱ_1	0.3	0.4	0.5	0.6	0.7	0.8	0.9
			$\xi_1 = 0.2, \xi_2 = 0.2, \xi_3 = 0.4, \xi_4 = 0.2$				
0.3						1.5	3.9
0.4					1.7	4.2	6.4
0.5				1.8	4.4	6.7	8.7
0.6			1.8	4.5	6.9	8.9	10.7
0.7		1.7	4.6	7.0	9.1	10.9	12.6
0.8	1.7	4.6	7.1	9.3	11.2	12.8	14.3
0.9	4.5	7.1	9.4	11.3	13.0	14.5	15.9
			$\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.2, \xi_4 = 0.4$				
0.3						4.5	9.1
0.4					4.6	9.3	13.6
0.5				4.8	9.6	14.0	17.9
0.6			4.9	9.9	14.3	18.3	22.0
0.7		5.1	10.1	14.7	18.8	22.5	25.9
0.8	5.2	10.4	15.1	19.3	23.1	26.5	29.7
0.9	10.8	15.5	19.8	23.6	27.1	30.3	33.2

Table 2

Reduction in Cost for Constant Variances

When $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, and $\xi_1 = 0.2$, $\xi_2 = 0.3$, $\xi_4 = 0.4$.

	α					
ϱ	0.5	0.6	0.7	0.8	0.9	0.95
100	.227	.175	.132	.094	.059	.040
20	.304	.254	.200	.169	.127	.101
10	.367	.321	.279	.238	.193	.164
5	.462	.423	.387	.351	.308	.277
2	.661	.646	.634	.621	.599	.578
1	.876	.895	.918	.943	.971	.985

APPENDIX

Minimizing the variance (17) with respect to C^* with the assumption of 100% coverage by frame A and the equality of variances, the optimum solution for p' is obtained as

$$p'^2 = \frac{1 - \alpha}{\varrho' - \alpha}\left\{\frac{\sigma^2_{a(1)}(\xi_1 + \xi_2)}{\sigma^2_{ab(1)}(\xi_3 + \xi_4)}\right\}$$

with

$$\varrho' = \frac{C'_A}{C'_B}.$$

Using $N_{A1} = N_{a(1)} + N_{a(12)} + N_{ab(1)} + N_{ab(12)}$ and $N_{B1} = N_{ab(1)} + N_{ab(12)}$, ϱ' can be written as

$$\varrho' = \frac{C_1}{C_4}\alpha\frac{N_a(\xi_1 + \xi_2) + N_{ab}(\xi_3 + \xi_4)}{N_{ab}(\xi_3 + \xi_4)}$$

$$= \frac{C_1}{C_4}\alpha\left(\frac{1 - \alpha}{\alpha}\frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} + 1\right)$$

$$= \frac{\alpha}{K}\left(\frac{1}{\alpha_1^*} + 1\right)$$

where

$$\alpha_1^* = \frac{\alpha}{1 - \alpha}\frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

Then we have

$$p'^2 = \frac{1 - \alpha}{\frac{\alpha}{K}\left(\frac{1}{\alpha_1^*} + 1\right) - \alpha}\frac{\xi_1 + \xi_2}{\xi_3 + \xi_4}\Phi'_1$$

$$= \frac{K\Phi'_1}{1 + \alpha_1^*(1 - K)} \qquad \text{(A.1)}$$

Define

$$\lambda_1 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_2 = q^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_3 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_4 = q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}.$$

With the p' in (A.1), the optimum sample sizes will be

$$\frac{n^2_{AO}}{N_A} = \gamma' \frac{(N_{a(1)} + N_{a(12)})\sigma^2_{ab(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(12)}}{C'_A}$$

$$= \gamma' \frac{\lambda_3}{C'_A}$$

$$\frac{n^2_{BO}}{N_B} = \gamma' \frac{q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}}{C'_B} = \gamma' \frac{\lambda'_4}{C'_B}$$

with γ' determined wih respect to (18). From this we get

$$\frac{n_{BO}}{n_{AO}} = \frac{N_B}{N_A}\left(\frac{C_1 N_{A1}\lambda_4}{C_4 N_{B1}\lambda_3}\right)^{1/2}. \tag{A.2}$$

Also, the variances given by (5) and (17) at optimum sample sizes can be written as

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\lambda_1 + \frac{N_B}{n_B}\lambda_2 \tag{A.3}$$

$$V(\hat{Y}^{*(1)}) = \frac{N_A}{n_{AO}}\lambda_3 + \frac{N_B}{n_{BO}}\lambda_4.$$

Equating the above variances and using (A.2), we obtain expression for n_{AO} and n_{BO} in terms of n_A and n_B as follows:

$$\frac{n_{AO}}{N_A} = \frac{\lambda_3 + \left(\frac{C_4\lambda_3\lambda_4 N_{B1}}{C_1 N_{A1}}\right)^{1/2}}{\frac{N_A}{n_A}\lambda_1 + \frac{N_B}{n_B}\lambda_2}$$

and

$$\frac{n_{BO}}{N_B} = \frac{\lambda_4 + \left(\frac{C_1\lambda_3\lambda_4 N_{A1}}{C_4 N_{B1}}\right)^{1/2}}{\frac{N_A}{n_A}\lambda_1 + \frac{N_B}{n_B}\lambda_2}.$$

Using these relationships, the cost C^* may be obtained as

$$C^* = \frac{(\xi_1 + \xi_2)\left[\left\{C_1(1 + \alpha_1^*)(\Phi_1' + \alpha_1^* p'^2)\right\}^{1/2} + \alpha_1^*\left(C_4 q'^2\right)^{1/2}\right]^2}{\frac{1}{1 - \alpha}\left\{\frac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \frac{\alpha\alpha_1^* q^2}{n_B}\right\}}. \qquad \text{(A.4)}$$

REFERENCES

ARMSTRONG, B. (1979). Test for multiple frames sampling technique for agricultural survey: New Brunswick, 1978. *Survey Methodology*, 5, 178-199.

BOSECKER, R.R., and FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *American Statistical Association Proceedings of the Social Statistics Section*, 219-224.

HARTLEY, H.O. (1962). Multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 203-206.

HARTLEY, H.O. (1974). Multiple frame methodology and selected application. *Sankhya*, Series C, 36, 99-118.

LUND, R.E. (1968). Estimation in multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 282-288.

SERRURIER, D., and PHILLIPS, J. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, 2, 138-170.

VOGEL, F.A. (1975). Surveys with overlapping frames, problems in application. *American Statistical Association Proceedings of the Social Statistics Section*, 695-699.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 133–144
Statistics Canada

Seasonal Adjustment of Labour Force Series during Recession and Non-Recession Periods

ESTELA BEE DAGUM and MARIETTA MORRY[1]

ABSTRACT

This paper analyzes the revisions of eight seasonally adjusted labour force series during recession and non-recession periods. The four seasonal adjustment methods applied are X-11 and X-11-ARIMA using either concurrent or forecast seasonal factors. The series are seasonally adjusted with these four methodologies according to both a multiplicative and an additive decomposition model. The results indicate that the X-11-ARIMA concurrent adjustment yields the smallest revisions both during recession and non-recession periods regardless of the decomposition model used.

KEY WORDS: Survey; X-11; X-11-ARIMA; Concurrent adjustment; Recession/non-recession.

1. INTRODUCTION

Seasonality in some of the labour force series may be subject to abrupt changes due to dramatic variations in their composition during the various stages of the business cycle. An important example is total unemployment. In relatively prosperous years, it consists mainly of persons shifting jobs, new entrants to the labour market, workers from the primary sector (agriculture, forestry, fishing, trapping, etc.) and construction (in the winter), and students seeking jobs (in the summer). On the other hand, during recessions, the number of unemployed increases quickly and the newly unemployed are mainly regular workers from heavy industries and related activities characterized by seasonal variations of smaller amplitudes and seasonal patterns different from those in 'normal' years. This kind of shift was observed in Canada in 1981-1982, where the total unadjusted unemployment rose from 790,000 in August 1981 to 1,494,000 in December 1982; the newly unemployed coming mainly from the manufacturing and service industries.

The rapid changes in the size and composition of total unemployment during the depressed phase of the business-cycle raises the question as to whether the procedure followed to estimate seasonal factors based on data for years of low, mainly frictional and 'outdoor' unemployment, is applicable to data for years of high unemployment with a large number of the jobless added from the secondary and tertiary sectors.

Empirical research at Statistics Canada in 1974 led to current seasonal adjustment of labour force series by the X-11-ARIMA method using concurrent seasonal factors. This method of adjustment will be referred to as the 'official' procedure in the sections to follow. The U.S. Bureau of Labor Statistics officially adopted the X-11-ARIMA method in 1980 using six-month-ahead projected seasonal factors. This agency also releases monthly the unemployment rate calculated with X-11-ARIMA and concurrent seasonal factors. Concurrent seasonal factors are obtained by seasonally adjusting, each month, all the data available up to and including that month whereas projected seasonal factors are generated from data that ended usually one year before (in the case of the Bureau of Labor Statistics, six-months before).

In Section 2, the mean absolute error (MAE) of concurrent and year-ahead projected seasonal factors is given for eight Canadian labour force series obtained from X-11-ARIMA

[1] Estela Bee Dagum and Marietta Morry, Time Series Research and Analysis Division, Statistics Canada, 13th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

and X-11 using the multiplicative seasonal adjustment option. Year-ahead instead of six-months-ahead projected factors are analyzed because they are applied by several government statistical agencies. Furthermore, the MAE's of six-months-ahead factors fall between those of concurrent and year-ahead projected factors.

The main purpose of this study is to assess whether the use of X-11-ARIMA with concurrent seasonal factors still produces the smallest revisions during recession years when compared to three feasible alternative procedures.

In Section 3, the mean absolute revisions of the additive current seasonal adjustment are calculated for the four alternative procedures and MAE's of the additive are compareed to the multiplicative options.

Finally, the conclusions of this study are presented in Section 4.

2. REVISIONS OF CURRENT SEASONALLY ADJUSTED LABOUR FORCE SERIES DURING RECESSION AND NON-RECESSION PERIODS

The majority of the seasonal adjustment methods applied by government statistical agencies are based on linear smoothing filters, usually known as moving averages. It is inherent to these methods that the estimates from the observations of the most recent years are less accurate than those corresponding to central data because of the asymmetry of the end point filters. Among these methods, the Method II-X-11 variant developed by Shiskin, Young, and Musgrave (1967) and X-11-ARIMA developed by Dagum (1980) are the most widely applied. The X-11-ARIMA method is a modified version of the X-11 variant that basically consists of two steps. First, the original series are extended with extrapolation values from ARIMA (autoregressive integrated moving averages) models of the type developed by Box and Jenkins (1970), and then the extended series are seasonally adjusted with a set of moving averages that result from the combination of the X-11 seasonal filters with the extrapolation ARIMA filters. Therefore, the seasonal adjustment filters of X-11-ARIMA and X-11 differ for the data of the most recent year. For both procedures the same symmetric filter is applied to central observations. If the ARIMA option is not used, then the X-11-ARIMA reduces to the X-11 method.

As more data become available, the seasonally adjusted estimate pertaining to a time point keeps getting revised until the data point in question is three years away from the end of the series and the symmetric filters apply, at which point the estimate becomes virtually fixed and is referred to as the final seasonally adjusted estimate. The difference between the very first and the final seasonally adjusted estimate is called the total revision. The revisions of current seasonally adjusted values by the X-11-ARIMA and X-11 methods are due to: (1) Differences in the smoothing linear filters applied to the same observations as more data become available; and, (2) the innovations that enter into the series with new observations. One would like to see the revisions of the first kind reduced to a minimum or completely eliminated.

Theoretical studies by one of the authors (Dagum 1982a and 1982b) have shown that the revisions of current seasonally adjusted values due to filter changes can be reduced substantially if: (1) the original series is extended with ARIMA extrapolated values i.e., the X-11-ARIMA is applied; and (2) concurrent seasonal factors are used instead of year-ahead seasonal factors. The conclusion drawn from these two theoretical studies conforms to the results given in several empirical and theoretical works (see e.g. Dagum 1978, Dagum and Morry 1982, Kuiper 1978, 1981; Pierce 1980; Kenny and Durbin 1982; McKenzie 1982; Wallis 1982; Pierce and McKenzie 1985; Otto 1985).

Next, we examine the performance of X-11-ARIMA with concurrent seasonal factors compared to three other feasible alternatives for recession and non-recession periods. The better seasonal adjustment procedure will be the one that yields smaller revisions.

2.1 Comparisons of Four Alternative Procedures for Current Seasonal Adjustment of Labour Force Series

There are four seasonal adjustment procedures commonly applied to obtain current seasonally adjusted values, namely:

(1) X-11-ARIMA with concurrent seasonal factors;

(2) X-11 with concurrent seasonal factors;

(3) X-11-ARIMA with year-ahead projected seasonal factors; and

(4) X-11 with year-ahead projected seasonal factors.

The revision measure used here for the evaluation of the four alternative procedures is the mean absolute error (MAE) of the seasonal factors for current seasonal adjustment defined by:

$$MAE(N) = \sum_{t=1}^{N} \left| \hat{S}_t^c - \hat{S}_t^F \right| / N \qquad (1)$$

In this expression, N is the number of datapoints included in the mean, denotes the current seasonal factor value which can be either a concurrent or a year-ahead projected seasonal factor from X-11 or X-11-ARIMA. denotes the 'final' seasonal factor in the sense that it will not change significantly when the series is augmented with new data. For X-11 and X-11-ARIMA, a current seasonal factor becomes final when at least three years of data are added to the series (Young 1968; Wallis 1974). This study analyzes the revisions in the seasonal factors (or implicit seasonal factors in the additive case) rather than in the seasonally adjusted estimates for several reasons. First, using seasonal factors provides a feel for the size of revisions relative to the level of the series (it is in the form of a percentage); second, it standardizes the revision size within series subject to substantial jumps in level (such as the unemployment series); third, it allows for cross-series comparisons.

This study analyzes the revisions in the seasonal factors (or implicit seasonal factors in the additive case) rather than in the seasonally adjusted estimates for several reasons. First, using seasonal factors provides a feel for the size of revisions relative to the level of the series (it is in the form of a percentage); second, it standardizes the revision size within series subject to substantial jumps in level (such as the unemployment series); third, it allows for cross-series comparisons.

Unlike in a previous paper by the authors (Dagum and Morry 1982), the revisions in the month-to-month movement of the seasonally adjusted data were not included in the analysis since these revisions are not of primary interest when dealing with labour force data (for example, Statistics Canada does not publish yearly revisions of the growth-rate for these series). Consequently, this paper focuses on the revisions in the level rather than on revisions in the change in level.

The eight Canadian series of employment and unemployment analyzed here start in January 1966 and end in October 1982. To use the ARIMA extrapolation option of X-11-ARIMA a period of at least five years is necessary to produce a seasonally adjusted series. Consequently, the first year for which total revision measures can be calculated is 1971. Taking into account the need for at least three and a half more years for a current estimate to become final, the last full year for which MAE can be obtained is 1977. Within this seven-year span of revisions, we distinguished two years of recession and five years of non-recession. The recession period includes data from August 1974 until July 1975 and June 1976 until May 1977. These two years were considered recessionary because they showed high increases (greater than 25%) in the annual levels of total unemployment due mainly to large inflows of job losers.

Another important aspect taken into consideration is the kind of decomposition model used for the seasonal adjustment of each series. The X-11 and the X-11-ARIMA methods provide both additive and multiplicative decomposition models. There are no theoretical reasons for one model to be preferable to the other. They are based on different assumptions concerning the generating mechanism of the seasonal component.

In an additive model, the components of a time series (trend-cycle, seasonal variations and irregular fluctuations) are assumed to be independent and, therefore, the seasonal effect is not affected by the level of the economic activity conditioned by the stages of the business cycle.

On the other hand, in a multiplicative model, the seasonal effect is proportional to the trend-cycle. If the seasonal factors are constant, it means the higher the level of the seasonally adjusted series, the higher the seasonal effect.

The selection of the decomposition model is not crucial for the estimation of 'final' seasonally adjusted values since for most cases the corresponding figures are similar. The problem of model selection, however, becomes very important when approached from the viewpoint of the estimation of the seasonal component of the end years of a series, particularly, of series with a rapidly growing trend-cycle. The asymmetric filters used for the end points estimation, particularly those of the X-11 method, introduce large systematic errors if the seasonal estimates change fast (Dagum 1978). In fact, if the underlying decomposition model is that of a rather stable multiplicative seasonality, an additive seasonal adjustment will produce seasonal estimates that appear to vary with the trend-cycle. Reciprocally, if stable additive seasonality is the norm, a multiplicative adjustment will produce seasonal factors that look unstable or fast moving.

From the viewpoint of seasonal adjustment, it is then preferable to choose the decomposition model that yields the most stable seasonal estimates. The tests developed by Morry (1975) and Higginson (1977) have been applied to the eight series to determine the preferred decomposition models.

The results of these tests indicated that only two series, unemployment of adult and young women, follow an additive model; the remaining series are of the multiplicative type.

In this study, however, the mean absolute revisions have been analyzed under both assumptions, that is, the components of each series are either multiplicatively or additively related. We are using additive and multiplicative decomposition models for data spanning both recessionary and non-recessionary periods in order to determine which of these two decomposition models is more sensitive to sudden changes of level from the viewpoint of revision.

The calculations shown in the following tables were obtained from multiplicative seasonal adjustment. The results from additive adjustment are discussed in Section 3.

Table 1 shows the mean absolute error (MAE) of the seasonal factors of X-11-ARIMA and X-11 applied for current seasonal adjustment during recession years. It is apparent that X-11- ARIMA with concurrent seasonal factors yields the smallest revisions. This result is consistent with the theoretical findings discussed above which determined that the use of the ARIMA extrapolation option with concurrent seasonal factors significantly reduces filter revisions.

For six out of the eight series analyzed, X-11 with concurrent seasonal factors ranks second. For the other two series (unemployed and employed adult men) X-11/concurrent shows the same MAE results as does X-11-ARIMA with year-ahead projected seasonal factors. Finally, the least accurate estimates are obtained from X-11 with year-ahead projected seasonal factors.

Table 2 shows the relative size of the revisions from each alternative procedure with respect to X-11-ARIMA with concurrent seasonal factors during recession years. All the values are greater than 1.0 indicating that none of the alternative options gives revisions smaller than X-11-ARIMA/concurrent.

Table 1

Mean Absolute Errors (MAE(N)) of Seasonal Factors of X-11-ARIMA and X-11 during Recession Years[a] (N = 24)

Series	Concurrent Seasonal Factors		Year-ahead Projected Seasonal Factors	
	X-11-ARIMA (1)	X-11 (2)	X-11-ARIMA (3)	X-11 (4)
Unemployment				
Men 25+	1.95	2.75	2.74	3.35
Women 25+	1.94	2.94	3.43	4.70
Men 15-24	2.16	3.02	3.49	4.33
Women 15-24	1.25	1.73	2.48	3.44
Employment				
Men 25+	0.08	0.12	0.12	0.16
Women 25+	0.23	0.29	0.33	0.42
Men 15-24	0.41	0.53	0.66	0.76
Women 15-24	0.50	0.70	0.81	0.97

[a] August 1974 - July 1975 and June 1976 - May 1977.

Table 2

Comparison of MAE(N)'s from Three Alternative Procedures Versus X-11-ARIMA/Concurrent for Multiplicative Seasonal Adjustment of Employment and Unemployment Series in Recession Years (N = 24)

Series	X-11 Concurrent vs. X-11-ARIMA Concurrent (1)[a]	X-11-ARIMA Projected Factors vs. X-11-ARIMA Concurrent (2)[b]	X-11 Projected Factors vs. X-11-ARIMA Concurrent (3)[c]
Unemployment			
Men 25+	1.41	1.40	1.72
Women 25+	1.52	1.77	2.41
Men 15-24	1.40	1.61	2.00
Women 15-24	1.38	1.98	2.75
Employment			
Men 25+	1.50	1.50	1.50
Women 25+	1.26	1.43	1.83
Men 15-24	1.29	1.61	1.85
Women 15-24	1.40	1.62	1.94

[a] (1) equals column (2) ÷ column (1) of Table 1.
[b] (2) equals column (3) ÷ column (1) of Table 1.
[c] (3) equals column (4) ÷ column (1) of Table 1.

The non-recession period covers from January 1971 to December 1977 excluding the recession years. Table 3 shows the MAE of the current seasonally adjusted series for the four procedures during these years. Similarly to Table 1, X-11-ARIMA with concurrent seasonal factors yields the smallest revisions for all the series due to minimal filter revisions as pointed out before. For seven out of the eight series X-11/concurrent ranks second with values relatively close to those shown for X-11-ARIMA with year-ahead projected factors. Finally, the most unreliable procedure in terms of the magnitude of the revision is X-11 with year-ahead seasonal factors.

The relative size of the revisions of the three alternative procedures with respect to the X-11-ARIMA/concurrent procedure during non-recession years are shown in Table 4. The figures in column (1) with the exception of one entry, however, are smaller than those shown in column (1) of Table 2 which would indicate that during recession years the percentage gains achieved by using ARIMA extrapolation are even higher than during non-recession years.

Finally, Table 5 compares the size of the revisions during recession versus non-recession years for the two best procedures. The results show that X-11-ARIMA/concurrent which is Statistics Canada official procedure gives smaller MAE values compared to those of the second best alternative, X-11/concurrent. Most of the ratios in the first column are very close to 1.0, indicating that the revisions in times of recession are similar in size to those in non-recession years when using the ARIMA extrapolation option. If X-11 with concurrent seasonal factors is applied, the size of revision is substantially higher in most series during recession than in 'normal' times. This is due to the fact that the rapid change in the level of the series, introduced by the new observations of the recession years, is not estimated as well by the end filters. In fact, gradual movements and some of the level increase are passed to the seasonal component.

Table 3

Mean Absolute Errors (MAE(N)) of Seasonal Factors of X-11-ARIMA and X-11 during Recession Years[a] ($N = 60$)

Series	Concurrent Seasonal Factors		Year-ahead Projected Seasonal Factors	
	X-11-ARIMA (1)	X-11 (2)	X-11-ARIMA (3)	X-11 (4)
Unemployment				
Men 25+	1.37	1.73	2.22	2.73
Women 25+	1.84	2.41	2.92	3.55
Men 15-24	1.97	2.66	3.17	3.96
Women 15-24	1.93	2.87	2.59	3.18
Employment				
Men 25+	0.08	0.10	0.12	0.13
Women 25+	0.23	0.27	0.33	0.34
Men 15-24	0.39	0.46	0.58	0.69
Women 15-24	0.43	0.49	0.68	0.80

[a] From January 1971 until December 1977 excluding recession periods defined in Table 1 footnote (*a*)

Table 4

Comparison of MAE(N)'s from Three Alternative Procedures Versus X-11-ARIMA/Concurrent for Multiplicative Seasonal Adjustment of Employment and Unemployment Series in Recession Years (N = 60)

Series	X-11 Concurrent vs. X-11-ARIMA Concurrent (1)[a]	X-11-ARIMA Projected Factors vs. X-11-ARIMA Concurrent (2)[b]	X-11 Projected Factors vs. X-11-ARIMA Concurrent (3)[c]
Unemployment			
Men 25+	1.26	1.62	1.99
Women 25+	1.31	1.59	1.93
Men 15-24	1.35	1.61	2.01
Women 15-24	1.49	1.34	1.65
Employment			
Men 25+	1.25	1.50	1.62
Women 25+	1.17	1.43	1.48
Men 15-24	1.18	1.49	1.77
Women 15-24	1.14	1.58	1.86

[a] (1) equals column (2) ÷ column (1) of Table 3.
[b] (2) equals column (3) ÷ column (1) of Table 3.
[c] (3) equals column (4) ÷ column (1) of Table 3.

Table 5

Comparison of MAE(N)'s of Concurrent Seasonal Factors of X-11-ARIMA and X-11 for Recession Versus Non-Recession Years Using the Multiplicative Option

Series	X-11-ARIMA Concurrent Recession Years (N = 24) vs. Non-Recession Years (N = 60) (1)[a]	X-11 Concurrent Recession Years (N = 24) vs. Non-Recession Years (N = 60) (2)[b]
Unemployment		
Men 25+	1.42	1.59
Women 25+	1.05	1.22
Men 15-24	1.09	1,35
Women 15-24	0.67	0.60
Employment		
Men 25+	1.00	1.20
Women 25+	1.00	1.07
Men 15-24	1.05	1.27
Women 15-24	1.16	1.54

[a] (1) equal to column (1) of Table 1 ÷ column (1) of Table 3.
[b] (2) equal to column (2) of Table 1 ÷ column (2) of Table 3.

The only exception is the series unemployed women 15 to 24 where revisions with both methods are smaller during economic hardship. This can be explained by the special behaviour of this series during the period analyzed, which is characterized by large annual increases of about 15% for 1966-73 and 8.5% for 1973-80 and an additive seasonal component, independent of the business-cycle (i.e., the change in level reflected more the changing behaviour of young women than the effect of the business-cycle).

Another special case is the series unemployed men 25 years and over. Here recession years were characterized by much larger revisions than non-recession periods even with ARIMA extrapolations as indicated by a ratio of 1.42. This large discrepancy between the two periods is a result of the drastic composition changes in seasonality that this series undergoes during times of recession as discussed before. Without ARIMA extrapolation, the revision sizes deviate even more (the ratio is 1.59), since apart from the changes in composition the unreliable seasonal estimates produced during recession introduce added discrepancies.

3. COMPARISON OF ADDITIVE VERSUS MULTIPLICATIVE CURRENT SEASONAL ADJUSTMENT DURING RECESSION AND NON-RECESSION PERIODS

It is often argued that during recession periods the use of an additive instead of a multiplicative decomposition model is to be preferred from the viewpoint of the minimization of revisions. The main reasons given for this are: (1) in an additive model, the time series components are assumed to be independent and, therefore, the seasonal effect is not affected by the level of the trend-cycle contrary to what occurs with a multiplicative model; and (2) the inflexibility of the end-point filters to estimate adequately fast-moving seasonality.

The eight labour force series analyzed in the previous section was additively seasonally adjusted in order to assess this new alternative. The results obtained confirm the ranking given by the multiplicative option. Namely, X-11- ARIMA/concurrent yields the smallest revisions followed by X- 11/concurrent and X-11-ARIMA/year-ahead projected, in that order. The least accurate estimates are obtained with X-11/year-ahead projected. It is important to note that *factors* of additive seasonal adjustment mean *implicit* factors in the sense that they result from the quotient between the original series and the seasonally adjusted series.

Tables 6 and 7 show the relative size of the revisions by each alternative procedure with respect to X-11-ARIMA/concurrent, for the recession and non-recession periods, respectively. All the values are greater than one indicating that none of the alternative procedures gives smaller revisions than X-11- ARIMA/concurrent. Since the latter ranks first for both additive and multiplicative seasonal adjustment options, we compare for each series which of the two decomposition models gives the smallest revisions.

In Table 8 the data show that for the two series that affect the unemployment rate the most, i.e., the unemployment and employment of adult men, the multiplicative option is to be preferred during recession as well as non-recession years. For the most part, these data confirm the decomposition models chosen by Statistics Canada according to the model tests (Morry 1975; Higginson 1977). The only apparent exception is the series Employed Men 15-24 which would do better with an additive model. However, given the fact that the size of the revisions is already very small, this improvement is of no consequence. The MAE's from the multiplicative adjustment are 0.41 (recession period) and 0.39 (non-recession period) and are reduced by the additive options to 0.33 and 0.31 respectively.

Finally, we observe that the unemployment of adult women would have smaller revisions with a multiplicative instead of an additive seasonal adjustment during recession years.

Table 6

Comparison of MAE(N)'s from Three Alternative Procedures Versus X-11-ARIMA/Cncurrent for Additive Seasonal Adjustment of Employment and Unemployment Series in Recession Years (N = 24)

Series	X-11 Concurrent	X-11-ARIMA Projected Implicit Factors	X-11 Projected Implicit Factors
	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent
Unemployment			
Men 25+	1.18	1.29	1.38
Women 25+	1.16	1.49	1.75
Men 15-24	1.21	1.48	1.70
Women 15-24	1.33	1.74	1.84
Employment			
Men 25+	1.44	1.69	2.08
Women 25+	1.26	1.33	1.65
Men 15-24	1.02	1.05	1.34
Women 15-24	1.50	1.50	2.05

Table 7

Comparison of MAE(N)'s from Three Alternative Procedures Versus X-11-ARIMA/Concurrent for Additive Seasonal Adjustment of Employment and Unemployment Series in Recession Years (N = 60)

Series	X-11 Concurrent	X-11-ARIMA Projected Implicit Factors	X-11 Projected Implicit Factors
	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent
Unemployment			
Men 25+	1.31	1.65	1.88
Women 25+	1.20	1.59	1.71
Men 15-24	1.22	1.57	1.89
Women 15-24	1.05	1.20	1.26
Employment			
Men 25+	1.16	1.24	1.54
Women 25+	1.10	1.27	1.30
Men 15-24	1.22	1.31	1.55
Women 15-24	1.41	1.68	2.16

Table 8

Comparison of MAE(N)'s of Seasonal Factors from Additive Versus Multiplicative X-11-ARIMA (Concurrent) Seasonal Adjustment during Recession and Non-Recession Periods

Series	(N = 24) Recession Period	(N = 60) Non-recession Period
	Additive X-11-ARIMA Concurrent / Multiplicative X-11-ARIMA Concurrent	Additive X-11-ARIMA Concurrent / Multiplicative X-11-ARIMA Concurrent
Unemployment		
Men 25+	1.25	1.15
Women 25+	1.14	0.88
Men 15-24	1.23	1.05
Women 15-24	0.93	0.85
Employment		
Men 25+	1.25	1.25
Women 25+	1.00	1.00
Men 15-24	0.80	0.80
Women 15-24	1.14	1.17

4. CONCLUSIONS

The results of Sections 2 and 3 can be summarized as follows:

(1) The X-11-ARIMA method with concurrent seasonal factors gives the smallest revisions for each series, whether an additive or a multiplicative seasonal adjustment is made, during both recession and non-recession years.

(2) The comparisons of the magnitude of the revision from additive versus multiplicative seasonal adjustment with X-II-ARIMA/concurrent indicate clearly that the two series that affect the unemployment rate most, unemployment and employment of adult men, are of the multiplicative type during times of recession as well as non-recession.

(3) During recession years, the use of X-11-ARIMA with year- ahead factors and of X-11/concurrent yields equal MAE's for employment and unemployment adult men. For the six remaining series, however, X-11/concurrent is the second best alternative.

(4) The least accurate current seasonal adjustment estimates for all series in all the situations discussed are obtained with X-11 with year-ahead projected seasonal factors.

(5) The comparisons of the revisions during recession versus non-recession periods from X-11-ARIMA/concurrent show that they are of relatively similar magnitude with the important exception being Unemployed Men 25 years and over, where revisions are much higher in recession years. This concurs with the fact that this series undergoes abrupt seasonal changes because of drastic variations in its composition. The larger revisions are mainly due to these new innovations.

On the other hand, the use of concurrent seasonal factors with X-11 shows, for most series, large discrepancies in the size of the revisions of these two periods. This is an indication that revisions result mainly from the inadequacy of the end filters to estimate well the rapidly changing levels of recession periods.

For only one series, Unemployed Women 15-24 years, the two best procedures yield revisions substantially larger in non-recessions compared to recessions. This can be explained by the special behaviour of this series during the analyzed period which is characterized by large annual increases of about 15% for 1966-73 and 8.5% for 1973-80, obscuring the effect of the business-cycle; and, a seasonal component independent of the business-cycle.

Given the above observations, we can feel confident that the official seasonal adjustment procedure at Statistics Canada will give best estimates among the alternatives considered during recession.

ACKNOWLEDGEMENT

The authors are thankful to two anonymous referees whose helpful suggestions contributed to the improvement of this paper.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden Day.

DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labor Force Series.* Washington, D.C.: U.S. Government Printing Office.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method.* Catalogue 12-564E, Ottawa, Canada: Statistics Canada.

DAGUM, E.B. (1982a). Revision of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.

DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.

DAGUM, E.B., and MORRY, M. (1982). The estimation of seasonal variations in consumer price indexes. *Proceedings of the Conference on "The Measurement of Prices"*, Catalogue 22-24, Ottawa, Canada: Statistics Canada.

HIGGINSON, J. (1977). Users manual for the decomposition model test. Research Paper No. 77-01-001, Seasonal Adjustment and Time Series Staff, Statistics Canada.

KENNY, P., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society*, Series A, 145, 1-41.

KUIPER, J. (1978). A survey and comparative analysis of various methods of seasonal adjustment. *Seasonal Analysis of Economic Time Series* (Ed. Arnold Zellner), Washington, D.C.: U.S. Government Printing Office, 59-76.

KUIPER, J. (1981): The treatment of extreme values in the X-11-ARIMA program. *Time Series Analysis and Forecasting*, (Eds. Anderson, O., and Perryman, M.R.), Amsterdam: North-Holland Publishing Co., 257-266.

McKENZIE, S. (1982). An evaluation of concurrent adjustment on Census Bureau time series. *Proceedings of the Business and Economics Section of the American Statistical Association.*

MORRY, M. (1975). A test for model selection. Research Paper. No. 75-12-016, Seasonal Adjustment and Time Series Staff, Statistics Canada.

OTTO, M. (1985). Effects of forecasts on the revisions of seasonally adjusted values using the X-11 seasonal adjustment procedure. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association* (forthcoming).

PIERCE, D. (1980). Data revision with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.

PIERCE, D., and McKENZIE, S. (1985). On concurrent seasonal adjustment. Technical Paper, U.S. Bureau of the Census.

SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 variant of census method II seasonal adjustment program. Technical Paper No. 15, U.S. Bureau of Census.

WALLIS, K.F. (1974). Seasonal adjustment and relations between variables. *Journal of the American Statistical Association*, 69, 18-31.

WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11 method. *Journal of the Royal Statistical Society*, Series A, 145, 74-85.

YOUNG, A.H. (1968). Linear approximations to census and BLS seasonal adjustment methods. *Journal of the American Statistical Association*, 63, 445-457.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 145–159
Statistics Canada

Relational Patterns between Total Unemployment and Unemployment Insurance Beneficiaries in Canada

ESTELA BEE DAGUM, GUY HUOT, NAZIRA GAIT, and NORMAND LANIEL[1]

ABSTRACT

This study purports to assess whether there are temporal relationships between Unemployment Insurance Beneficiaries, Total Unemployment, Job Losers and Job Leavers in Canada using univariate and multivariate time series methods. The results indicate that during 1975-82 the Unemployment Insurance Beneficiaries series leads: (1) Total Unemployment by one month and (2) Job Leavers by two months. On the other hand, there are evidence of a feedback relationship between Unemployment Insurance Beneficiaries and Job Losers.

KEY WORDS: Job losers; Job leavers; ARIMA; VARMA; Multivariate time series.

1. INTRODUCTION

Unemployment Insurance (UI) plays a key role in helping the national labour markets adjust to trade and demand-induced changes in production and employment patterns. The main function of UI as part of labour market policy is to provide adequate financial protection during temporary unemployment, to facilitate adjustments. By removing the immediate threat from unemployment, UI relieves job seekers of the need to yield to economic pressures by accepting jobs unsuited to their skills or abilities. It permits a more systematic or wide-ranging job search contributing to the efficient reallocation of human resources. Furthermore, when there are temporary plant layoffs, the objective of UI is met by providing income protection to laid-off workers, so the employer keeps an experienced labour force intact. This saves him/her the cost of recruiting and training new employees after a layoff. It also saves the employee from going through extreme dislocation to prevent financial hardship.

In any situation, UI must have enough flexibility to take into account prevailing economic circumstances which may limit the availability of other jobs and extended jobseekers' unemployment. In the Canadian UI program, this flexibility is provided as longer benefit durations are triggered by rising regional unemployment rates.

The gap between overall unemployment and the UI series tends to narrow in recession and widen in recovery periods. Where business conditions worsen and layoffs occur, job losers become a greater proportion of Total Unemployment. As the most Unemployment Insurance claimants are in fact job losers, this increases the proportion of Unemployment Insurance Beneficiaries related to Total Unemployment.

This study purports to assess whether there is a temporal relationship between the Unemployment Insurance Beneficiaries and Total Unemployment in Canada. The analysis is extended to Job Losers (JLo) and Job Leavers (JLe) who can claim for benefits and are the two major groups of Total Unemployment. The existence of strong relationships among these variables can be useful to explain labour markets behaviour. Furthermore, they may lead to other types of similar relationships useful to estimate unemployment in small areas

[1] E.B. Dagum and G. Huot, Time Series Research and Analysis Division, Statistics Canada. N. Gait, University of Sao Paulo, Brazil, was visiting Statistics Canada when the paper was written, and N. Laniel, previously Time Series Research and Analysis Division, currently with Business Survey Methods Division, Statistics Canada.

where the sample size of the current labour force survey is inadequate. Section 2 introduces the definition of each of the four series discussed and analyzes the main characteristics from their spectra. Section 3 estimates the residual cross-correlation values, for several time lags, of the whitened series to assess whether or not there are pairwise relationships and their direction, if present. The residuals are computed from ARIMA models fitted to each series. Section 4 extends the previous analyses by identifying and estimating two multivariate time series models in order to understand the joint dynamic relationships of: (1) UIB and TU; and (2) UIB, JLo and JLe. Finally, Section 5 gives the main conclusions of this study.

2. THE MAIN CHARACTERISTICS OF THE ANALYZED SERIES

To understand the type of relationship between UIB and TU and its major components, JLo and JLe, we first introduce the definitions and analyze the main characteristics looking at their spectra.

2.1 Total Unemployment (TU)

The Labour Force Survey (LFS) Division of Statistics Canada obtains monthly information through a sample of 56,000 representative households across the country. Although developed since 1952, substantial revisions were introduced to the LFS from 1976.

Estimates of employment, unemployment and non-labour force activity refer to the specific week covered by the survey each month, normally the week containing the 15th day. The sample is designed to represent all persons in the population 15 years of age and over, residing in Canada, with some minor expections.

The Labour Force is composed to people who, during the reference week, were employed or unemployed. The employed includes persons who:

- did any work at all;
- had a job but were not at work due to illness or disability, bad weather, labour dispute, vacation, personal or family responsibilities.

The unemployed includes persons who:

- were without work, but actively looked for work in the past four weeks and were available for work;
- had not actively looked for work in the past four weeks but had been on layoff for 26 weeks or less, and were available for work;
- had not actively looked for work in the past four weeks but had a new job to start in four weeks or less, and were available for work.

Total unemployment is composed of the sum of job losers (JLo), job leavers (JLe), new entrants to the labour market, re-entrants after one year or less, re-entrants after more than one year (Statistics Canada 1976). Of these five components, the first two are the most important for our study since they can claim benefits and represent about 70% of TU.

Data on the flows into unemployment are not available prior to 1975. Thus, all the series were observed for the period January 1975 to December 1982, thus including the most recent data available at the time.

Figure 1 shows the original Total Unemployment series which is characterized by a peak in the winter months and a trough in the summer. Figure 2 shows the spectrum of the Total Unemployment series. High power is observed at the frequency 0.05 cycle/month associated with the business-cycle (0.05 corresponds to a 20-months cycle). Similarly, relatively high power is observed at the fundamental seasonal frequency 0.083 cycle/month and neighbouring frequencies, but less at the harmonics of the fundamental seasonal. Finally, the contribution of the irregular fluctuations to the total variance is small, relative to the other two components.

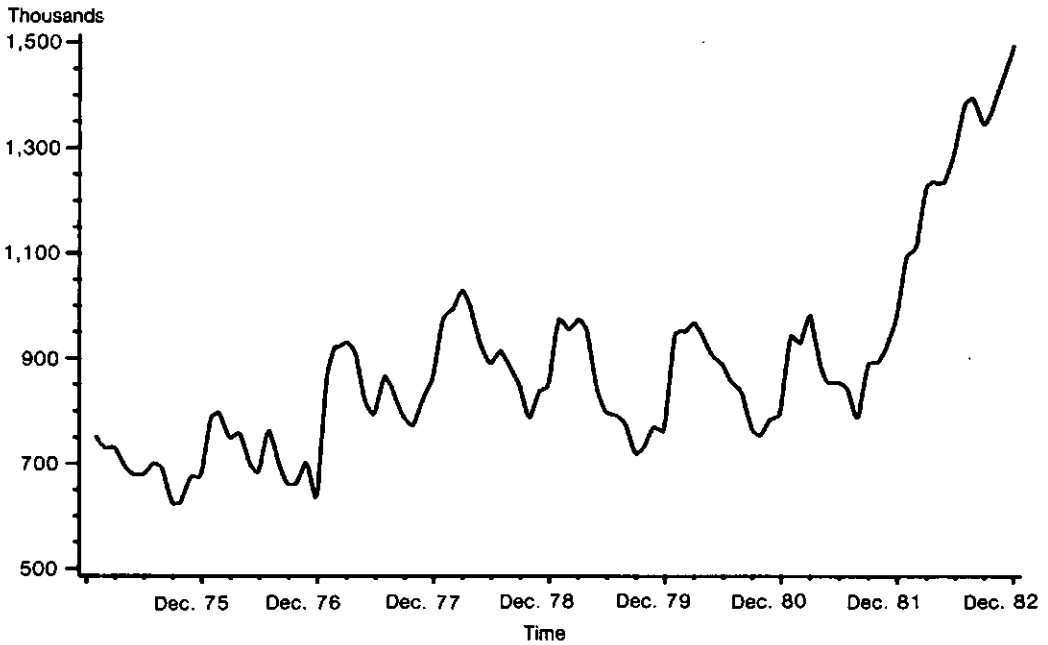


Figure 1. Total Unemployment Series

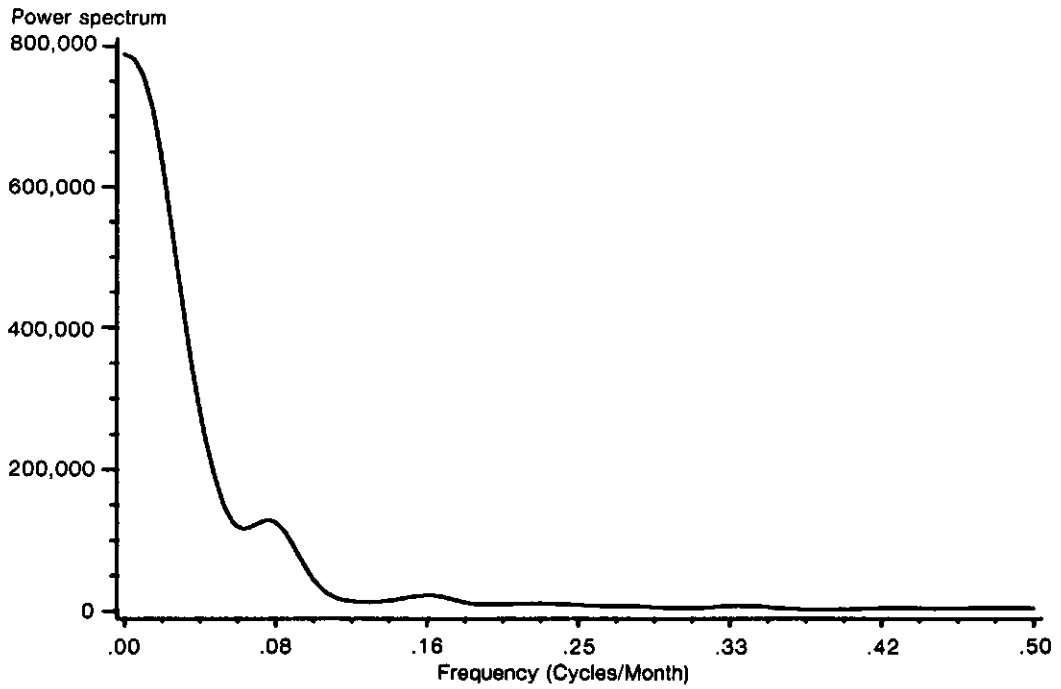


Figure 2. Spectrum of Total Unemployment

Figure 3 shows the original Job Losers series and Figure 4 displays its corresponding spectrum. Similar to TU, high power is shown at the business-cycle frequencies, but now most of the seasonal power is at the fundamental seasonal band and very little is left at the harmonic bands. The contribution of the irregular variations is smaller than that of TU.

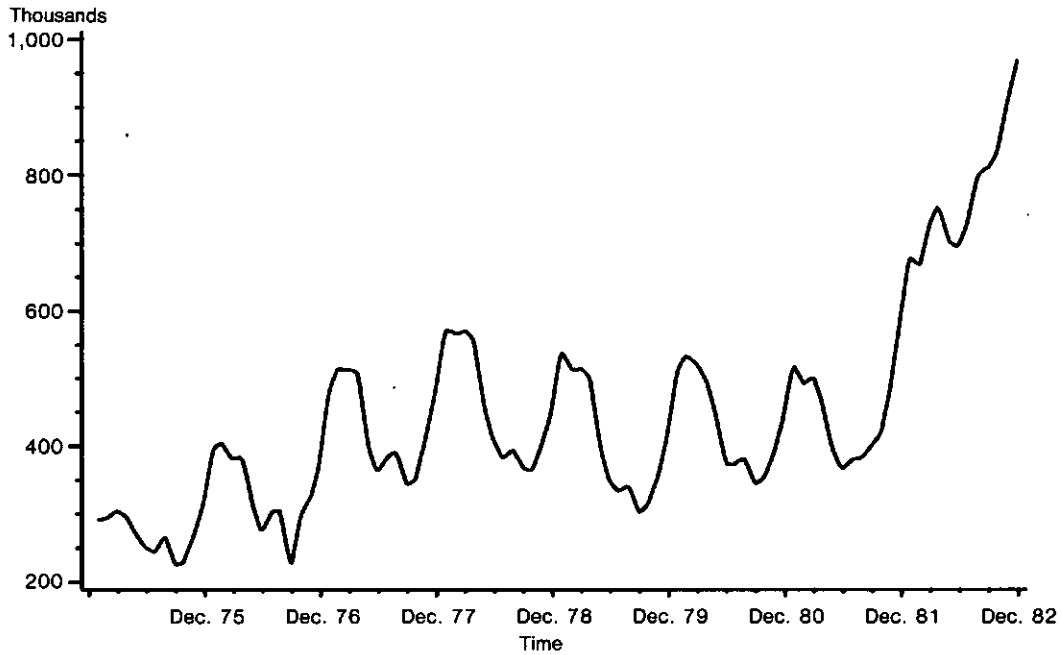


Figure 3. Job Losers Series

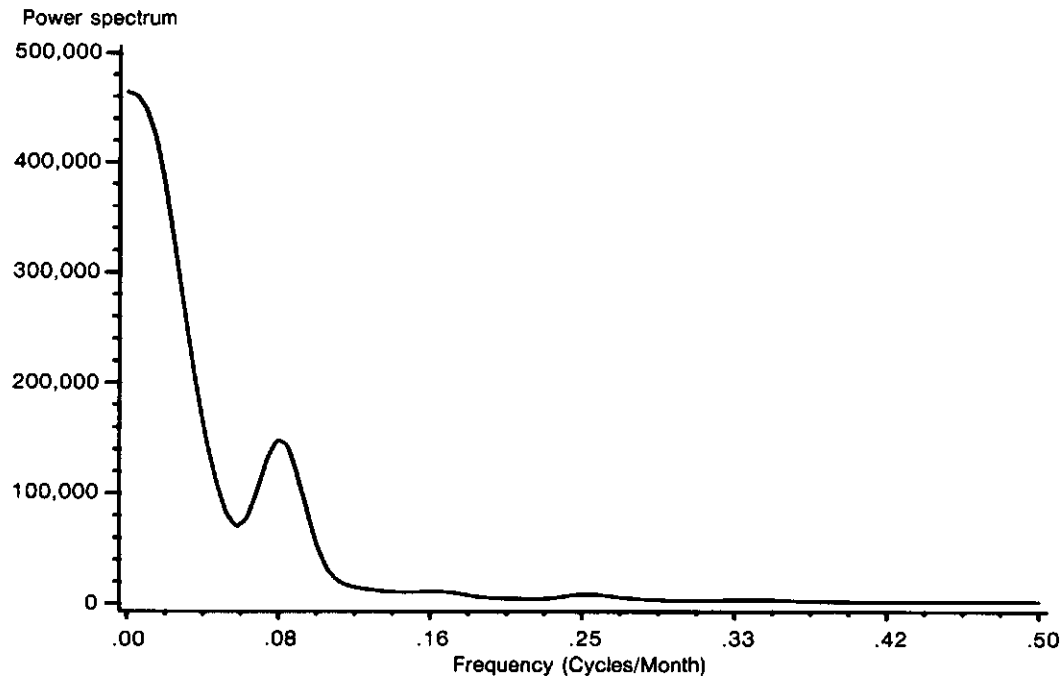


Figure 4. Spectrum of Job Losers

Figure 5 shows the Job Leavers series, characterized by two troughs, one in the winter months and the other during the summer. Its spectrum is given in Figure 6. This series has more cyclical variations than trend as indicated by the high peak at 0.022 cycle/month which corresponds to a 45 months-cycle. Furthermore, the seasonal variations are highly concentrated around the first harmonic band, supporting the fact that this series has two seasonal troughs. Finally, the contribution of the irregular to the total variance is larger than that observed for the two previous series.

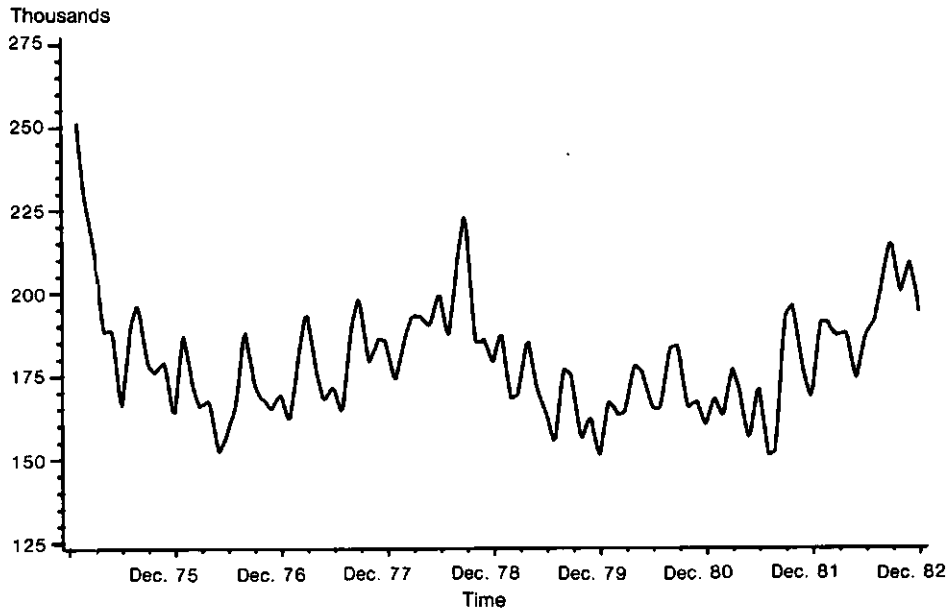


Figure 5. Job Leavers Series

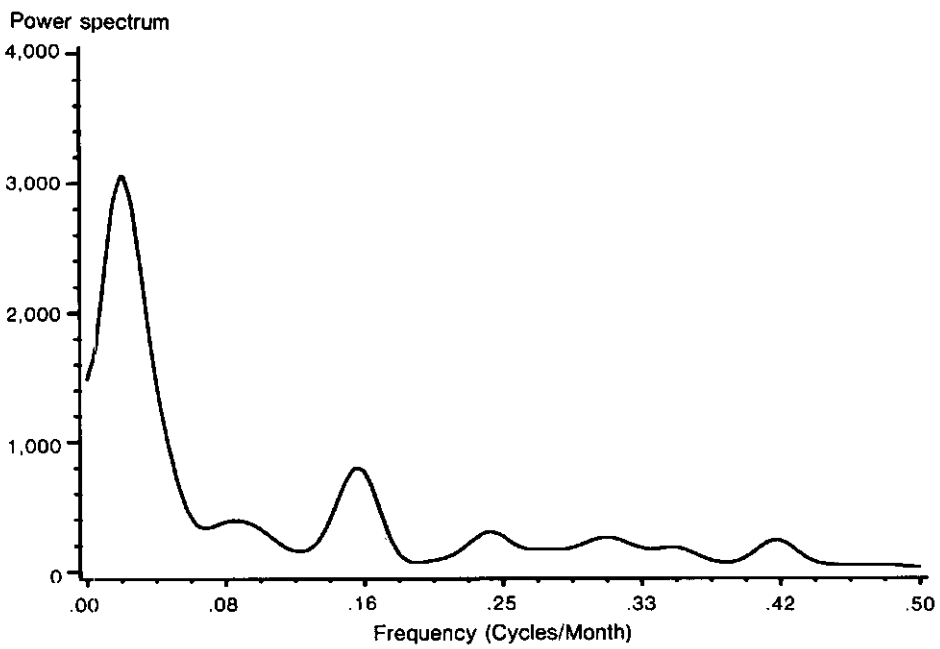


Figure 6. Spectum of Job Leavers

2.2 The Unemployment Insurance Beneficiaries (UIB)

The monthly data for Unemployment Insurance Beneficiaries cover all persons drawing benefits for a specific week, namely the week of the LFS. This is not a sample since it includes the total population of beneficiaries. The UI covers virtually all paid workers in the labour force and members of Armed Forces. The main exceptions are:

- People 65 years of age and over;
- People working fewer than 15 hours weekly;
- People earning less than 20% of the maximum weekly insurable earnings (in 1982, it was $70).

In order to qualify for benefits, a claimant must be available for and capable of work, unable to find suitable employment and have the necessary qualifying requirements. Previously eight weeks of work was the minimum required to qualify for benefits but as of December 1977 this number varied between 10 and 14 weeks according to the rate of unemployment prevailing in the region of residence of the claimant. Benefits are paid after a two-week period has been served.

Claimants who qualify for benefits can receive up to 25 percent of their benefits in earnings and continue to receive UI. However, the LFS considers these individuals to be employed. In order to assess the relationship between UI beneficiaries and unemployment, it is thus more accurate to use the series of UI beneficiaries *without* earnings. This subset of UI beneficiaries is a fairly consistent and significant proportion of the total LFS count of the unemployed. We must note, however, that because of differences in definition, the following groups are counted as unemployed in the LFS but are not included in the UI records, namely, entrants and re-entrants; all individuals who have worked but not long enough to qualify for benefits; and those unemployed persons who were previously self-employed. On the other hand, persons insured under the UI program can receive benefits even though, under the LFS definition they would not be classified as unemployed, examples include self-employed fishermen during the off-season, women on maternity leave and employees away from work due to sickness or disability.

The UI beneficiaries (without earnings) series is a sensitive indicator of labour market economic conditions. It is reflective of the insured labour force with recent work experience.

The original Unemployment Insurance Beneficiaries series, as shown in Figure 7, displays large seasonal fluctuations with a peak during the winter months, when bad weather curtails outdoor work in such industries such as fishing, construction and lumber, bringing a sharp rise in claims filed by affected workers.

Figure 8 shows the spectrum of the UIB series. Very high power is shown at the frequency 0.0167 cycle/month, which corresponds to a 60 months-cycle, and at those frequencies associated with the fundamental seasonal band. The contribution of seasonal variations to the total variance of the series is much larger than that observed in TU and its two major components. Finally, there is little irregularity relative to the trend-cycle and the seasonal components.

3. PAIRWISE RELATIONSHIPS BETWEEN UNEMPLOYMENT INSURANCE BENEFICIARIES, TOTAL UNEMPLOYMENT, JOB LOSERS AND JOB LEAVERS

Several early Canadian studies (e.g., Grubel *et al.* 1975; Green and Cousineau 1976; Jump and Rea 1975; and Siedule *et al.* 1976) support the general conclusion that unemployment has tended to shift upward with the increased availability of unemployment insurance in 1971. Lazar (1978) shows that the 1971 changes increased the unemployment duration and induced higher rates of job leaving, especially of young persons and adult women. These studies

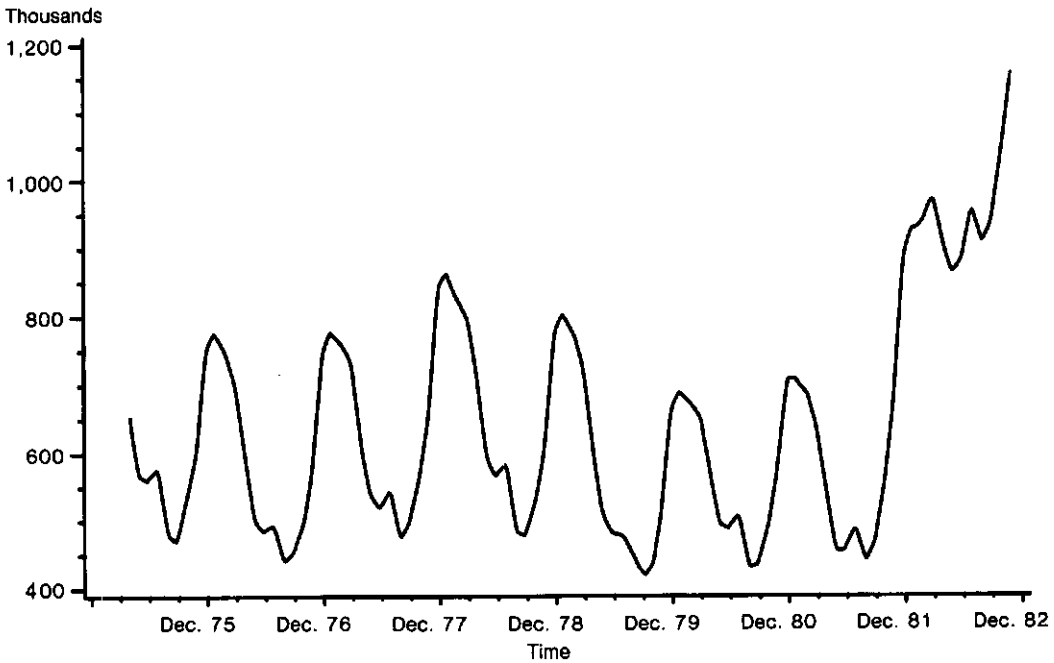


Figure 7. Unemployment Insurance Beneficiaries Series

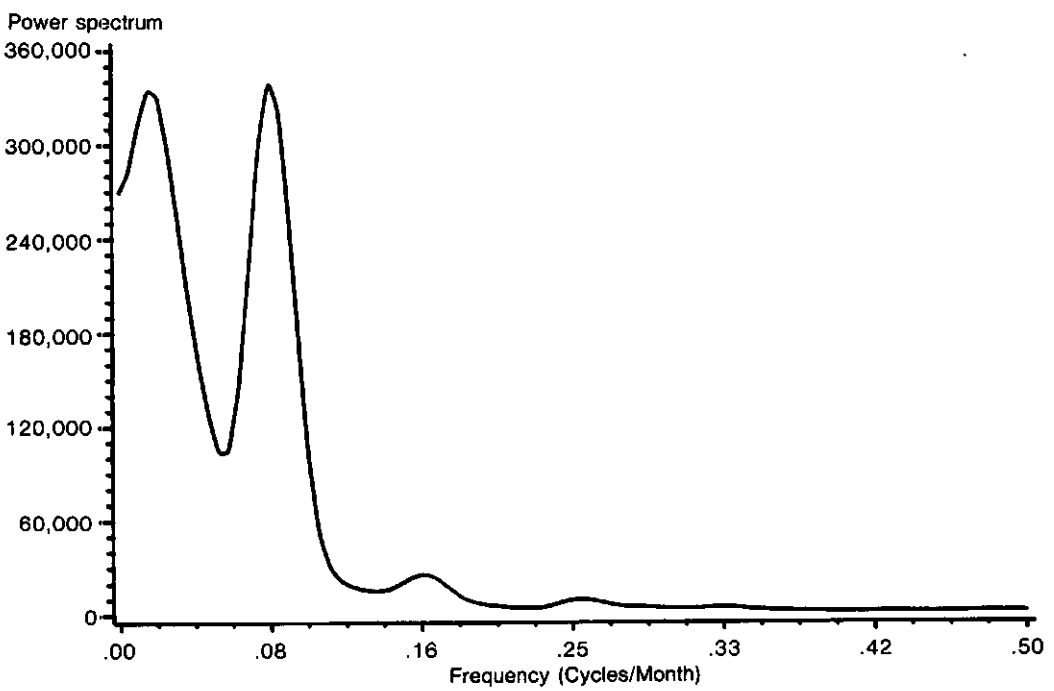


Figure 8. Spectrum of Unemployment Insurance Beneficiaries

were made before the changes of 1975 that aimed at strengthening work incentives. It was expected that the changes introduced after 1975 would reverse the effects of the program on total unemployment.

In this section, we carry out an exploratory analysis by searching for pairwise temporal relationships between Total Unemployment, Unemployment Insurance Beneficiaries, Job Losers and Job Leavers. The existence of these relationships will be useful to build a multivariate time series model to explain the joint dynamic behaviour of the above variables.

The pairwise relationships between TU, UIB, JLo and JLe are calculated using the cross-correlations of the residuals or *innovations* from ARIMA models (Box and Jenkins 1970) that fitted well the data. It has been rightly argued by several authors (e.g., Pierce and Haugh 1977) that the cross-correlations between white noise residuals obtained with different filters are biased to accepting the nul hypothesis of independence when it does not exist. Pierce and Haugh (1977) suggest to use dynamic regression models. This, however, implies that we have to make a judgement on which variable is the *cause* and which is the *effect*. At this stage, we are simply interested in determining whether there is a temporal relationship in each pair of variables analyzed. Table 1 shows the ARIMA models fitted to each series, their parameter values estimated with unconditional least squares, the results of the portmanteau test (Ljung and Box 1978) and the residual variance.

The Q statistics values accept the null hypothesis of randomness of the residuals in each case. However, since this test is applied to a set of autocorrelations of residuls for various lags, it is possible to have significant autocorrelation for some particular time lag k that will not be detected by this test. Therefore, we also tested whether there was autocorrelation of the residuals for each time lag. We used a more accurate approximation for small samples than $1/N$ to test the variance of the autocorrelation, that is, $(N - |k|)N^{-2}$ as given by Haugh (1976).

Having obtained satisfactory results from the above models we calculated the cross-correlation $\hat{r}_{xy}(k)$ between the series analyzed. The S_M^* statistic (Haugh 1976) is applied to test the independence between the series. Under the assumption that the residuals are normally distributed and that $E[\hat{r}_{xy}(k)] = 0$ and $\text{Var}\,[\hat{r}_{xy}(k)] = (N - |k|)N^{-2}$, the statistic

$$S_M^* = N^2 \sum_{k=-M}^{M} (N - |k|)^{-1} \hat{r}_{xy}(k)^2$$

follows a X^2 distribution with $2M + 1$ degrees of freedom. In order to determine the direction of the pairwise relationships, we modified the S_M^* statistics which is calculated for positive or negative k only, excluding zero.

Table 2 presents the estimates of the cross-correlation between Unemployment Insurance Beneficiaries (UIB) and Total Unemployment (TU) and its two major subcomponents Job Losers (JLo) and Job Leavers (JLe). We indicate with (a) and (b) those values significant at a 5% and 1% confidence level. In the case of UIB and JLo we calculated S_M^* for positive and negative values of k from ± 1 to ± 6 and from ± 1 to ± 2 to determine whether there is a dominant unidirectional relationship. The results indicated that there is no dominant direction between the two variables but a feedback process.

We can summarize the results from Table 2 as follows:

(1) There is indication of a unidirectional relationship between UIB and TU such that UIB would lead TU by one month;

(2) There is a feedback between UIB and JLo with a strong instantaneous relationship. Taking into consideration the time lag between the two variables, the feedback process seems to be initiated by JLo at lag 2.

Table 1
Univariate ARIMA Models

Series	ARIMA Models	Q(24)	$\hat{\sigma}_a^2$
Unemployment Insurance Beneficiaries (UIB)	$(1 - 0.68B)\Delta\Delta^{12} \log_{10} UIB_t = (1 - 0.80B^{12})a_t$	11.55	0.000140
Total Unemployment (TU)	$(1 - 0.25B^3)\Delta\Delta^{12} \log_{10} TU_t = (1 - 0.84B^{12})a_t$	9.13	0.000395
Job Losers (JLo)	$(1 - 0.31B^3)\Delta\Delta^{12} \log_{10} JLo_t = (1 - 0.67B^{12})a_t$	15.78	0.000604
Job Leavers (JLe)	$(1 - 0.37B^3)\Delta\Delta^{12} \log_{10} JLe_t = (1 - 0.40B - 0.25B^2)(1 - 0.87B^{12})a_t$	14.58	0.000627

Table 2
Cross-Correlation Between Unemployment Insurance Beneficiaries and Total Unemployment and Its Two Major Components, Job Losers and Job Leavers

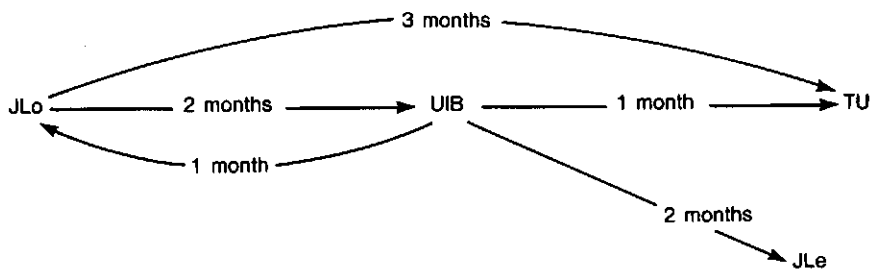
LAGS k	$UIB_{(t-k)} - TU_t$ $\hat{r}(k)$	$UIB_{(t-k)} - JLo_t$ $\hat{r}(k)$	$UIB_{(t-k)} - JLe_t$ $\hat{r}(k)$	$JLo_{(t-k)} - TU_t$ $\hat{r}(k)$
−6	−0.07	0.01	0.27[a]	0.15
−5	0.05	−0.04	−0.09	−0.04
−4	−0.09	0.03	−0.08	−0.01
−3	0.01	0.01	−0.11	−0.06
−2	0.14	0.28[a]	0.14	−0.01
−1	0.14	0.08	−0.01	0.21
0	0.16	0.32[b]	0.06	0.39[b]
1	0.22[a]	0.29[a]	0.04	0.14
2	0.12	0.12	0.26[a]	−0.16
3	−0.07	0.00	0.19	0.42[b]
4	0.12	−0.05	0.01	−0.05
5	−0.06	0.09	0.11	−0.04
6	0.13	0.00	0.08	0.05

[a] 5% significance level.
[b] 1% significance level.

(3) There is a unidirectional relationship between UIB and JLe such that UIB would lead JLe by 2 months. We observe, however, the effect of a delayed feedback at lag 6 which arises from the fact that the JLe series have a strong secondary peak in summer as shown in Figure 6.
(4) Finally, there is a strong instantaneous and undirectional relationship between JLo and TU such that JLo would lead TU by 3 months.

The above observations lead to the following Diagram 1 which will be useful for the identification of a more complex multivariate time series model that also takes into account the partial associations among the variables.

Diagram 1



4. BUILDING A MULTIVARIATE TIME SERIES MODEL FOR UNEMPLOYMENT INSURANCE BENEFICIARIES, TOTAL UNEMPLOYMENT, JOB LOSERS AND JOB LEAVERS

In the previous section we concluded that there are pairwise relationships among the four variables in the sense defined by Granger (1969) and Pierce and Haugh (1977). Taking into consideration those preliminary relationships, we here identify and estimate two multivariate time series models following the methodology developed by Tiao and Box (1981) and Tiao and Tsay (1983). These models will explain the joint dynamic behaviour of the variables involved.

A vector ARMA model for seasonal series takes the form

$$\underline{\phi}(B)\underline{\Phi}(B^s)\underline{Z}_t = \underline{\theta}(B)\underline{\Theta}(B^s)\underline{a}_t \tag{4.1}$$

where

$$\underline{\phi}(B) = \underline{I} - \underline{\phi}_1 B - \ldots - \underline{\phi}_p B^p \tag{4.2}$$

$$\underline{\Phi}(B^s) = \underline{I} - \underline{\Phi}_1 B^s - \ldots - \underline{\Phi}_P B^{sP} \tag{4.3}$$

$$\underline{\theta}(B) = \underline{I} - \underline{\theta}_1 B - \ldots - \underline{\theta}_q B^q \tag{4.4}$$

$$\underline{\Theta}(B^s) = \underline{I} - \underline{\Theta}_1 B^s - \ldots - \underline{\Theta}_Q B^{sQ} \tag{4.5}$$

are matrix polynomials in B (the back shift operator which is defined by $B^m Z_t = Z_{t-m}$), the $\underline{\phi}$'s, $\underline{\Phi}$'s, $\underline{\theta}$'s and $\underline{\Theta}$'s are $k \times k$ matrices, s is the seasonal periodicity and a_t is a sequence of random shock vectors *IID* $N(\underline{0}, \underline{\Sigma})$ and $\underline{Z}_t$ is a vector of stationary time series.

In order to avoid a problem of multicollinearity between TU and JLo, two VARMA models were specified, a VARMA $(1,2)(0,1)_{12}$ that relates Unemployment Insurance Beneficiaries with total Unemployment, and a VARMA $(2,6)(0,1)_{12}$ that relates UIB with Job Losers and Job Leavers. These models were identified and estimated using the exact maximum likelihood method in the Scientific Computing Associates program (Liu and Hudak 1983). The models are fitted respectively to the original data transformed as follows:

$$\begin{pmatrix} uib_t \\ tu_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} UIB_t \\ TU_t \end{pmatrix} \tag{4.6}$$

and,

$$\begin{pmatrix} uib_t \\ jlo_t \\ jle_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} UIB_t \\ JLo_t \\ JLe_t \end{pmatrix} \tag{4.7}$$

Table 3 shows the parameter values of the VARMA (1,2)(0,1) model and the standard errors of estimates given in parenthesis. (The estimated parameter values and the variance-covariance matrix of the residuals shown in Table 3 cannot be compared with the one of the univariate models (Table 1) because the former result from the fit of the model to the standardized transformed data instead of the non-standardized as it was the case with the univariate models.) Examination of the pattern of the cross-correlations of the residuals in Table 4 suggests that the model is adequate. A plus (minus) sign is used when the estimate is greater (less) than twice its standard error and a dot for a non-significant value based on the above criterion.

Thus, the VARMA model for UIB and TU becomes,

$$uib_t = 0.669 uib_{t-1} + \hat{a}_{1t} - 0.794\, \hat{a}_{1(t-12)} \tag{4.8}$$

$$\begin{aligned} tu_t = {} & 0.475 uib_{t-1} - 0.347 tu_{t-1} + \hat{a}_{2(t)} - 0.308 \hat{a}_{2(t-2)} \\ & - 0.705 \hat{a}_{2(t-12)} + 0.217 \hat{a}_{2(t-14)} \end{aligned} \tag{4.9}$$

Table 3

Estimated Parameters for the Transformed UIB and TU Variables

$\hat{\underline{\phi}}_1$

0.669	-
(0.089)	
0.475	−0.347
(0.098)	(0.115)

$\hat{\underline{\theta}}_2$

-	-
-	0.308
	(0.116)

$\hat{\underline{\Theta}}_{12}$

0.794	-
(0.090)	
-	0.705
	(0.086)

$\underline{\Sigma}$

0.429249	-
0.131532	0.544389

$\hat{\underline{\theta}}_1$

$\underline{0}$

Table 4
Cross-Correlation Matrices of the Residuals in Terms of +, −, and ·

LAGS 1 THROUGH 6

.. ..
.. ..

LAGS 7 THROUGH 12

.. ..
.. ..

LAGS 13 THROUGH 18

.. ..
.. ..

LAGS 19 THROUGH 24

.. ..
.. ..

Equations (4.8) and (4.9) indicate that Unemployment Beneficiaries leads the Total Unemployment series by one month. In fact, when analyzing the relationship between UIB and TU we must keep in mind that an increase in JLo and thus an increase in UIB may lead other members of the family to look for work in order to compensate for the loss of income. These are the new entrants and re-entrants who do not qualify for insurance benefits but contribute to an increase in TU. Furthermore, we should note that it is possible to have an increase in Total Unemployment without an increase in the normal gross flow of labour markets, simply because an increase in UIB occurs during recessionary periods where the availability of jobs is significantly reduced and thus flows into the unemployment state will increase.

The results of this model are in agreement with the preliminary results obtained from the pairwise cross-correlations of the previous section as shown in Diagram 1. The model, however, provides us with a more complete information on the dynamic behaviour of these two phenomena. We observe that the Unemployment Insurance Beneficiaries series is positively related to its previous-month level whereas the Total Unemployment is positively related to the previous-month level of UIB and negatively related to its previous-month level. In both equations, the effect of seasonality is reflected in their moving average part with a high parameter value for the random shock at lag 12.

Table 5 shows the VARMA $(2,6)(0,1)_{12}$ model applied to the transformed UIB, JLo and JLe variables as given in System (4.7).

Table 6 indicates no recognizable patterns in the estimated cross-correlation matrices of the residuals and, therefore, this model is considered adequate.

The final vector ARMA $(2,6))0,1)_{12}$ model for the three variables is,

$$uib_t = 0.617uib_{t-1} + 0.268jlo_{t-2} + \hat{a}_{1(t)} + 0.221\hat{a}_{3(t-6)} - 0.831\hat{a}_{1(t-12)} - 0.176\hat{a}_{3(t-18)} \tag{4.10}$$

$$jlo_t = 0.577uib_{t-1} - 0.285jlo_{t-1} + \hat{a}_{2(t)} + 0.386\hat{a}_{3(t-6)} - 0.525\hat{a}_{2(t-12)} - 0.308\hat{a}_{3(t-18)} \tag{4.11}$$

$$jle_t = 0.303uib_{t-2} - 0.411jle_{t-1} - 0.403jle_{t-2} + \hat{a}_{3(t)} - 0.797\hat{a}_{3(t-12)}. \tag{4.12}$$

Table 5

Estimated Parameters for the Transformed UIB, JLo and JLe Variables

$\hat{\underline{\Phi}}_1$

0.617 (0.086)	–	–
0.577 (0.099)	−0.285 (0.096)	–
–	–	−0.411 (0.088)

$\hat{\underline{\Phi}}_2$

–	0.268 (0.080)	–
–	–	–
0.303 (0.083)	–	−0.403 (0.084)

$\hat{\underline{\theta}}_6$

–	–	−0.221 (0.087)
–	–	−0.386 (0.108)
–	–	–

$\hat{\underline{\Theta}}_{12}$

0.831 (0.094)	–	–
–	0.525 (0.096)	–
–	–	0.797 (0.077)

$\underline{\Sigma}$

0.339	–	–
0.117	0.483	–
0.014	0.153	0.428

$\hat{\underline{\theta}}_t, t = 1,2,\ldots,5$

$\underline{0}$

Table 6

Cross-Correlation Matrices Terms of +, −, and ·

LAGS 1 THROUGH 6

...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

LAGS 7 THROUGH 12

..+	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

LAGS 13 THROUGH 18

.._	...	...	...	_..	...
...	...	...	...	...	...
...	...	...	...	...	...

LAGS 19 THROUGH 24

...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

Equation (4.10) and (4.11) shows the existence of feedback between Job Losers and Unemployment Insurance Beneficiaries similar to the relationship found in section 3. The JLo series leads UIB by two months (equation 4.10) and the one month lagged UIB strongly affects the current value of JLo (equation 4.11). Furthermore, each of the two endogeneous variables UIB and JLo are affected by their previous-month levels, positively in the case of UIB and negatively in the case of JLo. The relationship between both series due to seasonality is reflected by the parameter values of a_{t-6} and a_{t-12}. The need for a moving average term at lag 6 arises from the fact that the JLe series have a strong secondary peak in summer as shown in figure 6.

These empirical results are not in contradiction with economic theory. It has been argued, with good reason, that causality cannot be detected only from empirical evidences but must be supported by economic theory (see e.g. Zellner 1979). It is easy to accept that an increase in Job Losers which is associated with an economic recession will lead to an increase in Unemployment Insurance Beneficiaries. In turn, an increase in Unemployment Insurance Beneficiaries will lead to an increase in Job Losers because in reaction to a severe economic recession, most firms make temporary layoffs to be able to have their employees back when economic conditions improve.

Equation (4.12) raises an interesting question when showing that Unemployment Insurance Beneficiaries leads the Job Leavers by two months. It is not so evident why this should be the case.

Plausible explanations can be found in the analysis of the shortrun dynamics of the Canadian labour markets and a thorough investigation would require longitudinal data. We can, however, entertain the hypothesis among others that an increase in JLo and thus an increase in UIB may lead other members of the family to look for work in order to compensate for the loss of income. These persons are the new entrants and re-entrants. During a recession when JLo is increasing it is very difficult for new entrants and re-entrants to find a job. These new entrants and re-entrants are mainly young people and women over 25 who are willing to accept any job, at first, as long as it means extra income for the family. They might work for the length of time necessary for them to qualify for benefits. Then, once they qualify for benefits, they would become JLe in order to be more selective in the kind of job they will accept.

5. CONCLUSIONS

The main purpose of this study has been to assess whether there are temporal relationships between Unemployment Insurance Beneficiaries (UIB) and Total Unemployment (TU), Job Losers (JLo) and Job Leavers (JLe) by building dynamic multivariate time series models.

We have first carried out an exploratory analysis by searching for pairwise temporal relationships between TU, UIB, JLo and JLe in the sense defined by Granger (1969) and Pierce and Haugh (1977). Our results indicated the existence of relationships among the four variables involved.

We have then identified and estimated two multivariate time series models following the methodology developed by Tiao and Box (1981) and Tiao and Tsay (1983). The results of the vector ARMA models agree with the preliminary results obtained from the pairwise cross-correlations of the residuals of the univariate ARIMA models.

The first vector ARMA model shows that the UIB series leads TU by one month. UIB is also positively related to its previous-month level whereas TU is negatively related.

The second vector ARMA model shows that JLo leads UIB by two months with the existence of a one-month feedback from UIB to JLo. Furthermore, UIB is posivitely affected by its previous-month level while JLo is negatively related. It also shows that UIB leads JLe by two months.

These empirical results based on data for 1975-82 are not in contradiction with economic theory. Furthermore, they conform to those of earlier Canadian studies, based on data prior to 1975, which supported the general conclusions that the increased availability of unemployment insurance induced higher rates of job leaving, especially of young persons and adult women and led to increased levels of unemployment. Hence, it seems that the UIC regulation change in 1977 had little effect, if any, in this regard.

It would have been very interesting to assess the effect of the high recession that started in July 1981 but given the series length, elimination of this recessionary period would have made the series too short for any sound statistical modelling.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis Forecasting and Control*. San Francisco: Holden Day.

GRANGER, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral Methods. *Econometrica*, 37, 424-438.

GREEN, C., and COUSINEAU, J.M. (1976). Unemployment in Canada: The impact of unemployment insurance. *The Economic Council of Canada, Ottawa.*

GRUBEL, H.G., MAKI, D., and SAX, S. (1975). Real and insurance induced unemployment in Canada. *Canadian Journal of Economics*, VIII, 174-191.

HAUGH, L.D. (1976). Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of American Statistical Association*, 71, 378-385.

JUMP, G.V., and REA, S.A. (1975). The impact of the 1971 unemployment insurance act on work incentives and the aggregate labour market. *Institute for Policy Analysis*, University of Toronto.

LAZAR, F. (1978). The impact of the 1971 unemployment insurance revisions on unemployment rates: Another look. *Canadian Journal of Economics*, August, 559-570.

LIU, L.M., and HUDAK, G.B. (1983). *Univariate - Multivariate Time Series and General Statistical Analysis*. Illinois: Scientific Computing Associates.

LJUNG, G.M., and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.

PIERCE, D.A., and HAUGH, L.D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5, 265-293.

SIEDULE, T., SKOULAS, N., and NEWTON, K. (1976). The impact of economy-wide changes on the labour force, an econometric analysis. *The Economic Council of Canada*, Ottawa.

STATISTICS CANADA (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-716.

TIAO, G.C., and BOX, G.E.P. (1981). Modelling multiple time series with applications. *Journal of American Statistical Association*, 76, 802-816.

TIAO, G.C., and TSAY, R.S. (1983). Multiple time series modelling and extended sample cross-correlations. *Journal of Business and Economic Statistics*, 1, 43-56.

ZELLNER, A. (1979). Causality and econometrics. *Carnegie Rochester Conference Series on Public Policy*, Volume 10, (Eds. Karl Brunner and Allan H. Meltzer), Amsterdam: North-Holland Publishing Company, 9-54.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 161–170
Statistics Canada

Basic Principles of Questionnaire Design

LARRY SWAIN[1]

ABSTRACT

Thirty basic principles of questionnaire design are presented covering the content, wording, format, and testing of questionnaires. The extent to which the questionnaire is an integral part of the survey is emphasized as is consideration of its relationship with other aspects of survey design.

KEY WORDS: Survey; Questionnaire; Methodology.

1. INTRODUCTION

Most surveys make use of a questionnaire which is to be completed by either a respondent or an official representative of a survey organization (by personal contact or telephone). Since the questionnaire is the means by which the objectives of a survey are transformed into measurable variables, successful achievement of those objectives requires an effective questionnaire. In addition, the questionnaire may help structure, standardize, and control the data collection process so that the required information is obtained in a satisfactory manner. Effective questionnaire design is a combination of basic principles and common sense, adapted to the particular needs of each individual survey.

Although thirty separate principles of questionnaire design are presented, they are not intended to be seen as independent of each other or of the survey environment in which they operate. The extent to which the questionnaire is an integral part of the survey process cannot be sufficiently emphasized. As the questionnaire *cannot* be designed in isolation from the various other aspects of the survey, the reader is also advised to consider *during* questionnaire design its relationship with survey objectives, population, data collection, coding and data capture, editing, imputation, confidentiality, and testing.

Since this paper is not intended to be a comprehensive discussion of either survey or questionnaire design, alert readers, depending on their own perspectives of the various aspects of a survey, may identify omissions in the principles or may wish to exclude particular principles as more appropriate to a survey component other than questionnaire design.

The basic principles of questionnaire design as presented cover the content, wording, format, and testing of questionnaires. The questionnaire has a major impact on whether or not the survey objectives are met. Unlike other major survey components such as sample design or data processing procedures, the questionnaire directly involves the respondent. Therefore, it is essential that the content, wording, and format ensure the collection of reliable, valid, and relevant information from the respondent.

The author recognizes that although some of the principles appear obvious when stated, they are usually not so in practice. Also, some of the principles are measurable; some are not.

In the principles which follow, the term *questionnaire* is consistently used to refer to the various types of forms used to obtain information. In the literature and in practice, distinctions are often made among:

(a) a questionnaire (completed by a respondent);

---

[1] Larry Swain, formerly of the Census and Household Survey Methods Division, Statistics Canada; currently of the Human Resources Planning Division, Public Service Commission, Ottawa, Canada K1A 0M7.

(b) an interview schedule (completed by an interviewer);
(c) an administrative form (completed by a respondent or an official representative of the survey organization);
(d) a form used to record observations or measurements (completed by an official representative of the survey organization);
(e) a form used when transcribing information from existing administrative records (completed by an official representative of the survey organization).

For simplicity, the term *questionnaire* is used herein to represent all such forms. In addition, the term *questionnaire item* is used to represent the particular question or statement requesting information, including the response categories or space for response.

The term *survey* is used generally to represent any data collection activity, including sample surveys, censuses, and administrative data collection.

2. CONTENT

1. All questionnaire items should be directly related to the objectives and uses of the survey.

It is a reasonable goal that the collection of information be designed to minimize response burden by techniques such as reducing the number of questions. Exclusion of questionnaire items only remotely related to the objectives and uses of the survey is a means of satisfying this goal.

In addition, questionnaire items that ask for irrelevant information unnecessarily contribute to the overall length of a questionnaire and may provoke suspicion in respondents, factors which may lead to increased non-response rates (a possible source of bias), to a poorer quality of data because of fatigue or lack of concentration by interviewers or respondents, and to increased costs, both financial and temporal, to the survey sponsor and to the respondents.

For the questionnaire designer, the very act of relating each questionnaire item to the survey objectives and uses helps ensure that these objectives and uses are well defined and will indeed be satisfied by the questionnaire.

2. If a questionnaire contains items that, although relevant to the survey, may not appear so to respondents, then an explanation of the reason for their inclusion should be provided to respondents.

Classification variables such as age, sex, marital status, size of organization, number of employees and variables such as name, address, and telephone number (used for follow-up procedures or for editing purposes) are possible examples where an explanation to respondents should be considered for inclusion (at least at a general level).

3. Only those questionnaire items for which responses can be provided easily and with sufficient reliability should be included.

Where information is requested through recall by respondents, the events should be sufficiently recent or familiar to the respondents; where the request can be satisfied from available records maintained by respondents,the effort (including both time and cost) required to obtain the information should not exceed the benefits to be gained by acquisition of the information.

Because of potential definitional ambiguities, increased response bruden, and processing errors, it may be advisable that respondents not be asked to process information to complete a questionnaire item. It may be easier and more accurate for respondents to be asked for the specific information already available to them, to be processed later by the survey organization.

4. Respondents should not be asked questionnaire items for which they cannot be expected to provide any response.

Questionnaire items should not presume that the respondent has knowledge or awareness of a specific topic or engages in a particular activity. Filter questions can be used to exclude a respondent from a subsequent questionnaire item or sequence of items if those items are irrelevant because of the respondent's own particular characteristics, circumstances, or opinions.

Should respondents encounter many irrelevant items, they may feel that the survey questionnaire had been given to them in error. This could contribute to non-response or to poor relations with respondents.

The use of a filter question also serves to identify clearly whether or not a respondent is required to answer a subsequent questionnaire item or sequence of items. This is useful during survey processing and subsequent analysis. If a response to a questionnaire item is blank, it may be difficult to distinguish between the situation in which the reason for the blank is non-response (a refusal or an accidential omission), and that in which it is because the question does not apply (in the case of a numerical answer, the question may seem not to apply when the answer is legitimately zero). A filter question helps resolve this problem by identifying which respondents should have answered the questionnaire item.

Complex skip patterns, however, should be avoided, especially for those questionnaires completed by respondents themselves. Also, the number of filter questions should be minimized.

For items requiring a numerical answer, an alternative to a filter question is the inclusion of a *None* category.

3. WORDING

1. The phrasing of a questionnaire item should be appropriate to the respondent.

If a respondent does not understand a questionnaire item, it is probable that the response to that item will be inaccurate or not be given. Words, phraseology, and sentence structure familiar and appropriate to those providing the information should be used.

Abbreviations should be avoided unless they are understood by respondents.

2. Where there is sufficient demand, questionnaires should be translated into other languages.

Steps should be taken to ensure that the translated version corresponds adequately to the original version with respect to the intended meaning.

3. The questionnaire designer should choose the type(s) of questionnaire items most appropriate to obtain the required information while minimizing the response error and response burden in obtaining that information.

The types of questionnaire items for consideration are the open-response or free-answer type, the closed-response or fixed-answer type, and the fill-in-the-blanks type. Closed-response types are those items for which answer categories are provided. Fill-in-the-blanks types, although they appear to be open-response because no answer categories are explicitly provided,

are actually implicitly closed from the respondent's point of view in that the choice of answers is usually limited to a number, a day of the week, a province, etc.

Generally, closed-response questions entail less respondent and/or interviewer burden, since they do not require respondents to formulate and answer in their own words nor do the answers have to be recorded verbatim.

4. When a decision among two to more well-defined alternatives is required, a closed-response or fill-in-the-blanks type of questionnaire item should be used.

When all the alternatives are too numerous to be listed, then the use of the category *other* to represent a number of infrequently occurring responses, the use of an open-response or fill-in-the-blanks type of questionnaire item, or the collapsing of alternatives into fewer categories is recommended. In fact, it may be appropriate to use a fill-in-the-blanks type, where the response categories and numerical codes are included in a separate instruction booklet accompanying the questionnaire. In business, agriculture and institutional surveys, fill-in-the-blanks types of items are frequently used for questions that require a numerical response. The choice and number of categories in a closed-response type depends on the complexity of interpretation of the concept, the uses to which the data will be put, and the prior information available to the questionnaire designer.

5. When the alternatives to a question are not well-defined, an open-response type of questionnaire item should be used.

Open-response types of questionnaire items are frequently used in preliminary research or exploratory studies to generate specific hypotheses and to structure items for subsequent questionnaires. The open-response type of questionnaire item may also be used as a means of probing for additional or qualifying information, for purposes of verification of other questionnaire items, for use in interpreting data, as a change of pace, or as an introduction to a new topic.

6. If ease, timeliness, and cost of processing the data for capture are important considerations, closed-response types of questionnaire items should be used.

Open-response types of questionnaire items require coding of the information provided, an operation which can be both costly and time-consuming and is also subject to errors of interpretation and procedure.

In addition, with open-response types of questionnaire items, no specific frame of reference is provided, leading to the choice of varying frames of reference on the part of respondents. These varying frames of reference and the provision of varying amounts of information by respondents cause difficulty in the recording, coding and analysis of responses. On the other hand, a closed response provides a specific frame of reference, which although avoiding the above problems, may artificially induce a response. This is especially true when the repondent has little or no information or opinion about a particular topic. The questionnaire designer must therefore be aware of the possible frames of reference of respondents before choosing a type of questionnaire item.

Once the type and wording of a questionnaire item have been decided upon, restrictions are placed on the uses to which the information can be put, the specific hypotheses which can be tested and the analyses that will be applied to the item. This implies that the determination of objectives, uses, hypothesis testing and analyses is a prerequisite to the final version of the item. This determination does not preclude that the data will suggest additional analyses and uses within the limits imposed by the questionnaire items themselves.

In addition to the above considerations, the past experience of the questionnaire designer will contribute to the choice of suitable type(s) of questionnaire items in particular situations.

7. Response categories for closed-response types of questionnaire items should be non-overlapping and exhaustive (that is, mutually exclusive and comprehensive).

The response categories of a particular questionnaire item should be distinct and include all possibilities.

The distinctiveness of response categories does not preclude the applicability of more than one response to a particular questionnaire item. In such a case, a note such as *check as many as apply* should be included as part of the question.

Where response categories are such that only one response is to be provided to a particular questionnaire item, a note such as *check one item only* should be included as part of the question (except in the most obvious cases, for example, where the response categories are *Yes* and *No*). In those cases where more than one response can be applicable but where the designer wishes that only one item be checked in order to restrict responses, a note such as *check the most appropriate item* should be provided.

8. The units of response should be specified.

Either the units of response (e.g., kilograms, tons, per cent, hours per week) should be included in the questionnaire item or the respondent should be asked to specify them. Otherwise, there may be ambiguity as to which units were actually used.

9. Standardized concepts and definitions should be used.

To facilitate comparison of survey data with other sources of information (publications, other surveys) and to maximize the usefulness of the data (including secondary analysis), standardized (commonly understood and used) definitions should be used where they exist, and are well-defined, appropriate and up-to-date. Statistics Canada publishes standards related to occupational classes, industrial classes, commodities, geography and specific social concepts. In addition, Census concepts and categories are frequently used as standards.

10. The wording of questionnaire items should be specific, definitive, consistent, brief, simple and self-explanatory.

Survey concepts and terms that are new to repondents or subject to misinterpretation should be explained, defined, or avoided. To ensure consistent interpretation, the proper frame of reference (e.g., time reference, location, category of expenditure) should be provided. If consistency is required (e.g., different time references for different items), the change should be highlighted in the questionnaire.

Where several words can be used interchangeably, one of these should be selected and used throughout the questionnaire. If a synonym of a word already encountered is used in its place, respondents and others may assume that a different meaning is intended.

11. Double-barreled questions should be avoided.

A double-barreled question allows the respondent to make only one response although it is actually two questions in one. From the response, it is not possible to discern which

of the two ideas was answered or whether both were answered. The two issues should be asked separately except in specific circumstances where two issues necessarily have to asked together to convey the proper meaning. In such a case, it should be made clear to respondents that the two issues are both to be considered together.

12. Leading questionnaire items should be avoided.

A leading questionnaire item is one that is worded or formatted in such a way as to induce a respondent to choose a particular alternative or set of alternatives.

Some questions can be considered to be leading if they present options that may be perceived by respondents as socially unacceptable without an assurance that the respondent is made to feel that there would be no stigma attached to their response.

In attitudinal surveys, two basic principles have evolved to reduce (but not necessarily eliminate) response bias. The distribution of alternative answers should balance to provide approximately as many positive answers as negative to avoid leading respondents in one direction. Secondly, where there exists a series of items that have the same response alternatives, the sequence of items should either contain a mixture of positive and negative statements, be broken up, or be presented in a varied order to reduce the incidence of respondents answering in the same manner throughout the sequence (even though it may be inappropriate), without thinking very carefully about the particular responses.

4. FORMAT

1. Every questionnaire or questionnaire package should contain explanatory introductory material.
2. The introductory material should state the title of the survey, the name(s) of the sponsoring institution(s), and the purposes(s) of the survey.
3. An assurance to respondents of the confidentiality of the data that they provide should be considered.
4. The name (if appropriate) and telephone number or postal address of a contact within the sponsoring institution(s) should be included on the questionnaire in order that respondents may obtain additional information related to the survey, should they require it.

Generally, the introductory material may be in the form of a letter or brochure sent to the respondent; it might be a prepared statement made by an interviewer; or it can appear on the questionnaire itself. The introduction contains essential background information to respondents for the purposes of identification, legitimacy and notification of legal rights (if applicable).

5. Suitable identification should appear on the questionnaire.

For the purposes of estimation, field control, linkage with other records or follow-up on non-respondents, appropriate identification (numerical or otherwise) should be included on the questionnaire.

6. Questionnaire items and pages should be numbered.

To facilitate administration by interviewers, completion by respondents, and coding operations and instructions, questionnaire items and pages should be numbered consecutively (using either letters or numbers) throughout the questionnaire. If questions are written on

both sides of a page, an instruction (e.g., *over*) should appear at the bottom of the first side to ensure that the questions on the second side are completed.

7. The print on the questionnaire should be such that it can be easily read by the average respondent.

The person completing the questionnaire must be considered when determining the size of the type face (for example, small print could cause problems for those with poor eyesight) and the colours and contrasts of paper and type to be used. It is usually advisable to have different type face (size or type of characters) used for questions and instructions so that they can be easily distinguished.

8. Instructions for completion should be included on or with the questionnaire.

To help ensure that the questionnaire is completed properly by respondents, interviewers or other officials, brief but clear instructions should appear on or with the questionnaire (e.g., in an interviewers' manual or an instruction manual). However, questionnaire items should be as self-explanatory as possible to avoid complex sets of instructions.

For questionnaires being read by an optical character reader, clear instructions should be provided to help ensure their proper completion.

Instructions to respondents or interviewers for skipping items following filter questions should be sufficiently obvious and easy to follow. The use of arrows and directions may be appropriate. Complex skip patterns should be avoided, especially for questionnaires completed by respondents themselves.

9. The instructions for return procedures should be included on the questionnaire.

For a questionnaire which is to be returned by mail, the name and address of the person (or organization) to whom it is to be returned should be included on the questionnaire itself. Introductory letters and return envelopes can easily be mislaid or separated from the main body of the questionnaire.

The deadline by which respondents are to return completed questionnaires should also be stated.

For a questionnaire which is to be picked up by a field representative, space for the name and telephone number or postal address where the representative can be contacted and the date and approximate time of pick-up should be included on the questionnaire.

10. The numerical fields and codes used for data capture purposes should appear on the questionnaire (when capture is to be directly from the questionnaire).

When appropriate, data may be captured more quickly with fewer errors directly from the questionnaire itself. In such a case, the numerical fields and codes should be easily read by those performing the data capture but should not be a distraction to the respondent, interviewer or other official completing the questionnaire.

When data are to be coded before data capture, the coding boxes may appear on the questionnaire or on a separate sheet. When coding boxes do appear on the questionnaire, they should be clearly distinguished from answer boxes, perhaps with the *Office Use Only* designation or through appropriate shading.

Coding and data capture are often considered as steps that follow questionnaire design. It is essential for efficient implementation that they be considered during questionnaire design.

11. The format of answer spaces should be consistent throughout the questionnaire, with sufficent spacing for purposes of readability and accommodation of the responses to the questionnaire items.

Consistency of layout for response facilitates the task of a respondent, interviewer or official and aids in reducing error caused by inadvertent omission of a questionnaire item, an incorrect response, or a transposition of responses.

It may be useful to use different shapes for *check-off* type answers and numerical answers. One convention sometimes used is circles for the former and boxes for the latter.

There should be generous spacing on the questionnaire: to facilitate administration; to make the questionnaire more attractive and readable; and to provide the respondent, interviewer or official with sufficient space for the response to the questionnaire item.

12. Questionnaire items should be sequenced in a logical order for ease of completion and to provide the proper frame of reference.

The sequence of questionnaire items should appear logical to the respondent (a logic that may be different from that of the questionnaire designer), with questionnaire items related to one another grouped together. One sometimes recommended method is to have questions proceed from the most general questions to the most specific. Question ordering should try to anticipate the order in which respondents will supply information. The questionnaire designer should recognize that a question may prompt an answer not only to that question but also to another question which (hopefully) follows very shortly.

Transitions between sections of questions should be smooth. Section headings or introductory statements to sections should be used. For questionaires used in transcription from other documents, a logical sequence would be that of the source document.

In attitudinal surveys, the questionnaire designer should avoid conditioning respondents in the early questioning to a frame of reference which could bias responses to later questions. For example, questionnaire items regarding the awareness of a concept should precede any other mention of that concept. Sensitive questions should be placed within the context of related questions so as to justify their inclusion as much as possible and desensitize the questions somewhat.

13. The final version of the questionnaire should contain no typographical or grammatical errors.

The inclusion of errors on the questionnaire may have an adverse effect on data quality in that the questionnaire may not be treated seriously or may be misunderstood by those completing it. In addition, errors may contribute negatively to the image of the survey organization in the eyes of the public.

5. TESTING

1. Questionnaires administered for the first time or containing substantial modifications should be tested prior to their use as a collection document.

Just because all principles described in the previous principles have been followed, there is no guarantee that the proposed questionnaire will fully satisfy the objectives of the survey no matter how conscientious the researcher has been in designing the questionnaire. There are almost always unforeseen problems that occur in the administration of a questionnaire.

As a result, it is essential that a pretest of the questionnaire be implemented for all new surveys and for already existing surveys on which substantial modifications have been made in order to determine whether the objectives are likely to be met by the proposed questionnaire.

Some aspects of the questionnaire that the designer may test are the following: the wording, sequence and layout of the questionnaire to determine whether the questions and their flow are understood by respondents and interviewers; the necessity for inclusion of particular questions; the choice of types of questions; the use of specialized questioning techniques such as ranking or rating questions; the structure and definition of response categories; the degree of usage of the "other" category in questions; the ease of administration of the questionnaire; the time to administer various sections of the questionnaire; translation of the questionnaire; the possibility of bias in the questions; the nature of ethnic, regional or linguistic differences; the reasonableness of the questionnaire with respect to its demands on the respondent; the suitability of the questionnaire for measuring the concepts on which measurement is required; letters of introduction or introductory procedures; and the suitability of the method of collection.

A pretest should be done on at least a small sample of respondents (usually twenty to thirty) from the target population. It is preferable that the respondents be selected from the various subpopulations of the target population where differences or problems are likely to occur. Possible variables for definition of the test subpopulations are geographic region, educational background, age, sex, language, size of firm and type of industry. Depending on the particular purposes of the pretest, either a probability or a non-probability sampling scheme may be required for the selection of respondents, although in most cases, the latter is employed. One possibility is to use a focus group discussion of the questionnaire as a part of the pretest procedure.

The method of collection used for the pretest should be identical to that planned for the main survey. However, a personal interview is recommended for at least a portion of the pretest respondents so that the interviewer can then record the respondents' reactions, both verbal and non-verbal, as well as their own suggestions and impressions. After each test interview, the interviewer can discuss difficulties that the respondent had, the interpretation of questions and response categories, and so on. These difficulties can then be discussed with the designer of the questionnaire, for example, in the context of a meeting among the questionnaire designer and the pretest interviewers to debrief them on the interviews. For some pretests, it may be preferable to use experienced, skilled interviewers in order to maximize the usefulness of the pretest.

The pretest is an often-neglected procedure. It will almost always suggest improvements or will at least give the designer some assurance that the questionnaire used in the main survey, a much more expensive proposition, will likely proceed fairly efficiently. Of course, there is never any guarantee that all problems will be solved, but most major ones should be. A pretest need not be expensive and need not require a great deal of time for implementation and is recommended for all new or modified questionnaires.

ACKNOWLEDGEMENTS

The author wishes to thank not only the referee and editorial board for their helpful comments but also the many persons who responded to an earlier distribution as part of a partial draft of "Survey Design Standards and Guidelines".

REFERENCES

ANDERSON, J.F., and BERDIE, D.R. (1974). *Questionnaires: Design and Use.* Metuchen, N.J.: The Scarecrow Press, Inc.

BERTHIER, N., and F. (1971). *Le sondage d'opinion.* Paris: Bordas.

BON, F. (1974). *Les sondages - peuvent-ils se tromper?* France: Calmann-Lévy.

CARSON, E. (1974). Questionnaire Design, Some Principles and Related Topics. Unpublished Manuscript, Statistics Canada.

CORBIN, R., SWAIN, L., and WILHELM, E. (1977). Exposé pour un atelier sur la conception des questionnaires. Unpublished manuscript. Statistics Canada.

CORBIN, R., SWAIN, L., and WILHELM, E. (1977). Outline for a Workshop on Basic Questionnaire Design. Unpublished manuscript, Statistics Canada.

GHIGLIONE, R., and MATALON, B. (1978). *Les enquêtes sociologiques: théories et pratique.* Paris: Colin.

JAVEAU, C. (1974). *L'enquête par questionnaire. Manuel à l'usage du praticien,* third edition. Bruxelles: Institut de Sociologie de l'Université Libre de Bruxelles.

MOSER, C.A., and KALTON, G.J. (1972). *Survey Methods in Social Investigation,* second edition. New York: Basic Books.

OPPENHEIM, A.N. (1966). *Questionnaire Design and Attitude Measurement.* London: Heinemann.

PAYNE, S.L. (1951). *The Art of Asking Questions.* Princeton: Princeton University Press.

STATISTICS CANADA (1979). Basic Questionnaire Design, second edition.Unpublished manuscript, Statistics Canada.

STATISTICS CANADA (1981). Conception des questionnaires, Manuel d'atelier, third edition. Unpublished manuscript, Statistics Canada.

WARWICK, D.P., and LININGER, C.A. (1975). *The Sample Survey: Theory and Practice.* New York: McGraw-Hill.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 171–179
Statistics Canada

An Overview of the Strengths and Weaknesses of the Selected Administrative Data Files[1]

RAVI B.P. VERMA and PIERRE PARENT[2]

ABSTRACT

Twelve administrative data files are reviewed to determine if some of them could be used to derive migration data, in case the universality of the currently used family allowance files be limited, as a result of federal legislation.

It is found that none of the twelve files have strengths and weaknesses strictly comparable to those of the family allowance files. Further developments of the Health Care, and to a lesser extent the Old Age Security files are highly recommended.

KEY WORDS: Administrative files; migration; qualitative evaluation.

1. INTRODUCTION

In Canada, both family allowance and income tax files have a wide range of utility in producing the migration and population estimates for the different geographic areas (see Statistics Canada Catalogue Nos. 91-001, 91-210, 91-211 and 91-212). Data from the family allowance files are made available within 2 to 3 months after the reference date. In contrast, income tax data are available within 12 to 15 months after the reference date. However, income tax data provide the estimates of migration flows for the census divisions, and also by age and sex.

In terms of accuracy of population estimates, both family allowance and income tax files are good and they are comparable (see Norris and Standish 1983; Norris 1983; Verma *et al.* 1984; Verma and Basavarajappa 1985). One of the special features of the family allowance and income tax files is the fact that they are national in character. Another feature is that the records contain addresses with the postal codes. Thus, this could provide the migration information for local areas. However, in recent years, there seems to be some possibility that family allowance could cease to be universal as a result of government legislation. For example, coverage might be limited to the lower- and middle-sectors of the population. If this file ceased to be universal, its utility as a migration data source would be very severely limited. Hence, our population estimation activities would be jeopardized which in turn would affect other programs as revenue sharing, involving the annual distribution of $20 billion among provinces.

For this reason, alternate sources need to be explored. An attempt is made here to assess the strengths and weaknesses of some of the selected administrative data files for estimating migration and population for provinces and territories, census divisions, census metropolitan areas and other regions in Canada.

The twelve administrative files are qualitatively evaluated as an alternative to family allowance files. On the basis of their strengths and weaknesses, they are divided into the following three groups:

---

[1] Abridged version of the paper presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa, Canada.

[2] Ravi B.P. Verma and Pierre Parent, Demography Division, Census and Demographic Statistics Branch, Statistics Canada, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Major potential files for estimating migration flows

i) Health Insurance Files
ii) Old Age Security File

Major potential files used as a symptomatic indicator of population change and net migration

iii) Hydro Connections
iv) Telephone Customers
v) School Enrollments

Other files with limited or uncertain potential for estimating migration flow/net migration

vi) Driver's License
vii Building Permits
viii) Unemployment Insurance Beneficiaries
ix) Labor Force Survey
x) Voters' List
xi) Retail Sales
xii) Trucking Statistics

1.1 Criteria for Evaluating Administrative Data Files

The assessment of the usefulness of the various administrative data sources for estimating interprovincial and intraprovincial migration is done with respect to ten criteria: universe, coverage, method of determining migration information, types of migration, characteristics of records, reference date/period (and monthly availability), time-lag, historical availability, consistency and computerization (Almond 1982).

The new data source would have high potential if it contains features of the family allowance files, as described in Table 1. The most important criteria are: coverage, timeliness, consistency, monthly or quarterly availability, disaggregation using the postal code or other geocodes. The file or set of files that can meet these standards would probably qualify as replacement source to family allowance.

2. MAJOR POTENTIAL FILES FOR ESTIMATING MIGRATION FLOWS

Health Insurance and Old Age Security files are major potential files for estimating migration flows among provinces, territories and census divisions. Strengths and weaknesses of each of these two files are presented below.

2.1 Health Insurance File

Health Insurance is a provincial responsibility. Each province thus keeps a file of people eligible for the program. All residents in the province (including newly arrived immigrants and foreign students) are covered by the provincial insurance, except for RCMP and Armed Forces personel, and for the federal penitentiary inmates, covered by the federal government. Everybody who establishes its residence in a province must fill out a proper application, from which data on in-migrants, by province of origin, and on international immigrants can be compiled. Virtually complete coverage, monthly availability, minimal time lags and information usually detailed by age, sex and family composition of the migrants are the main strengths of the files. There should also be a very strong incentive for interprovincial migrants to apply to the program. Consequently, migration data should be reliable.

Table 1

Description of the Administrative Data Files Currently Used to Derive Migration Data in Canada

Criteria	F.A. Monthly Statistics Report	F.A. M0024 File
Universe	Children in payment of F.A.	Children entitled to F.A. (as opposed to "in payment")
Coverage	25% of total population in 1984. Virtually 100% of children aged 0-17	Similar to F.A. Monthly Statistics
Method for determining status	Compilation of change of address notices	Compilation of change of address notices
Types of migration	Interprovincial migration, by province of origin and destination	Similar to F.A. Monthly Statistics, plus international migration
Characteristics	Origin-destination. Age: total 0-17 only. Family size: refers to the number of children in family	Origin-destination. Age: year and month of birth. Language (E or F). Type of account (regular, foster, foreign or agency)
Reference date/period	Month: refers to the amount of information processed during that time	Month: refers to month of real migration
Time-lags	Data processed a given month is available at the end of that month and refers to migration of approximately two months earlier	Data released semi-annually. Contains information on last six months' migration. Available approximately 3 months after end of semi-annual version
Historical availability	January 1974 onwards for children migration data. From 1947 to 1973, only information on family migration was available	December 1977 to present
Consistency	OVER TIME: change in 1974. Slight, problems since 1980. AMONG PROVINCES: good	OVER TIME: generally good. Slight problems since 1982. AMONG PROVINCES: problems with Ont. since 1982. Slight problems with Nfld. and N.S. in 1983
Level of computerization	In provincial offices, yes. But data are sent to Health and Welfare Canada central office on print-outs	Yes, well developed

Note: F.A. is an abbreviation for Family Allowance.

Table 1

Description of the Administrative Data Files Currently Used to Derive Migration Data in Canada (Concluded)

Criteria	F55 Program	Revenue Canada File
Universe	Children entitled to F.A.	Tax filers (must have filed two consecutive years)
Coverage	Similar to F.A. Monthly Statistics	Filers matched two consecutive years total up to approximately 75% of population aged 18 and over
Method for determining status	Symptomatic indicator	Comparison of the return address of matched returns. Correction is brought for unmatched returns
Types of migration	Net migration	Intraprovincial, Interprovincial and International
Characteristics	Number of children by geographical area. Age	Origin-destination. Broad age-sex group
Reference date/period	Twice a year, as of June 1, and December 1 (refers to the number of children entitled to F.A. as of these dates)	Year: refers to the period between two consecutive filings, i.e. approximately the April-March period. Used as June-May data
Time-lags	Available approximately three months after reference date	Preliminary available 6-8 months after end of reference period. Final data, 10-12 months
Historical availability	December 1977 to present, with entitlement information. Available back to 1974 for children in payment	1966-67 to present
Consistency	Generally good	Changes in tax laws results in-change in coverage and in number of matched returns over time and provinces
Level of computerization	Yes, well developed	Yes, well developed

There are also, however, certain weaknesses. The fundamental limitation is that neither Ontario nor Quebec can provide migration data. In the latter case, however, new developments are promising, but for Ontario, nothing is expected. Unless a special source is derived for Ontario, this would compromise the high potential of this file. There could also be a consistency problem, since each province independently administers its file.

At the subprovincial level, migration could also be derived since the Provincial Health Care offices should be informed of any change of address. In the facts, however, all changes are not known.

Health Care files could also be used in regression estimates, especially in provinces that run periodic address checks to clean the file and count only the desired population.

2.2 Old Age Security Records

Health and Welfare Canada is responsible for the administration of the Old Age Security file. Canadian residents aged 65 and over who totalled a sufficient number of years of residence in the country are eligible. It represents approximately 10% of the total population. Coverage among eligible people is virtually universal. Also the financial incentive to report change of address is very strong. Another strength of the file is its timely availability. Information on people moving in a given month is compiled and received by Statistics Canada two or three months later. Finally, Old Age Security, being a federal program, provides comparable data for the provinces; even if the information is compiled by provincial regional offices, they all follow the same procedure.

The main shortcoming of this file for migration estimates purpose is the fact that it refers to a small portion of the population (varying from 7.3% in Alberta to 12.2% in P.E.I.), the elderly moreover showing a rather different migration pattern than the rest of the population. Unlike child migration, which can obviously be related to adult migration and then be blown up to estimate total migration, no similar efficient method could be developed to estimate total migration from the Old Age Security file. Although this could not be used as the main source for migration estimates, however, this file could provide a very interesting estimate of the elderly migration.

3. MAJOR FILES USED AS SYMPTOMATIC INDICATORS OF POPULATION CHANGE AND OF NET MIGRATION

Data from some administrative files could be useful for generating total population estimates. For example, School Enrolments, Hydro Connections or Telephone Residential Customers could be used in regression techniques as symptomatic indicators (see McRae 1985 for an application of Hydro Connections to population estimates). This method and the corresponding sources are generally used for producing small area population estimates, but if no other technique gives valuable estimates at the provincial level, these sources will be seriously considered.

3.1 Hydro Connections

Electric companies keep files of their customers. Information on the type of account (residential, commercial, farm, ...) and the address and postal code of the customer are available. Coverage of residential households is virtually complete. Sometimes there is only one file for the province, but sometimes 2 companies (Manitoba and Newfoundland) or even more (B.C. and Ontario) provide the electric facilities within the province. In most provinces data can be produced for the entire territory, as of any date and within a short time-lag, but for a few provinces it can be hard to get the data. The main weaknesses of the files are of two kinds. In addition to the previously cited problem there may also be slight inconsistencies due to the difference between provincial definitions of residential households (since

it responds to administrative criteria), and even within one province, if more than one company is involved. Nevertheless, Hydro Connections could be a very good source for population estimates. As a matter of fact, they were tested in British Columbia, where population estimates for municipalities and school districts were produced. The results were good. This method could also be tested and eventually be extended to provincial level estimates, if need be.

3.2 Telephone Companies

In Canada, telephone services are insured by 14 major telephone companies. Information on customers with residental lines (address and postal code) is available. The situation is roughly similar to that of the Hydro Connections files. Data can usually be obtained for specified dates within a rather small delay and the coverage is fairly high. Here again, more than one company may serve a given province, and also, a company may serve more than one province. Despite the fact no estimate based on Telephone files has been tested in Statistics Canada, it is felt that they have the potential to produce good results.

3.3 School Enrollments

Each provincial government maintains a computerized file on students enrolled in its primary and secondary school system, containing information on school addresses with the postal code and on the number of students, by age and grade. Information on the number of students refers to September 30 and is available between 4 and 10 months after the reference date, the time-lag varying by province. The coverage of students is also very good.

There are some weaknesses associated with this file. For example, its annual character plays against its use for producing quarterly estimates. Also its date of reference (September 30 instead of June 1), along with the up to 10 months delay is another handicap. At the subprovincial level, finally, it often can be observed that some students reside in a given administrative region, but go to school in a different one. This also could affect the quality of the estimates. It should be pointed out here that the school enrolment data, at one time, were used in Statistics Canada (and also by the U.S. Bureau of the Census, using a component method developed by them. See U.S. Bureau of the Census 1973, Chap. 23, p. 51); the deviations associated with that method were much higher than those with other methods. In case no other file could provide adequate population estimates, regression estimates with that file could produce acceptable results, at the provincial level at least.

4. FILES WITH LIMITED POTENTIAL

4.1 Driver's License

Each province maintains a file listing persons aged 15 (or 16, or 17) and over licensed to operate a motor vehicle. Using the provincial files,migration could be estimated in two ways: 1) compilation of changes of driver's address for estimating flows of migration; and 2) as a symptomatic indicator of the population change, through the variation of the number of people licensed in a given region. Currently, Ontario uses drivers' licenses to estimate intraprovincial migration, but very few other provinces could provide migration flow information, especially at the subprovincial level. In order to do so, it would require too much work and consultation with the provincial ministry. Despite the fact that drivers are forced by the law to report their change of address, not all do so, and no sufficiently detailed statistics are available.

The driver's license file could also be used in regression techniques. Data available at any specified date in many cases and short delays are positive points. However, coverage and consistency concerns might affect the quality of the data. For example, 83% of adults in Saskatchewan own a driver's license, as against 73% in Manitoba, 85% of males and 62% of females accounting for the latter province's average proportion. In addition, the poor, comparatively recent immigrants, and Indians and residents of remote communities in the

North have below average rates for holding licenses (Stock 1981, p. 44). For estimate purposes, it is often preferable to have a 100% coverage of a small subpopulation (e.g. children) than an 80-85% coverage of a large subpopulation (e.g. adults), especially if the coverage is selective with respect to migration. Although it does not necessarily make a file inappropriate for estimate purposes, it affects its potential.

4.2 Building Permits

Statistics Canada collects new building permits for cities and rural areas in Canada. On average, the coverage rates vary between the urban (98.5%) and the rural (62.5%) areas. The building permit data are available on a monthly basis at the census division level. These data could be also used as a symptomatic indicator of the population change. However, one of the weaknesses of the building permit data is the fact that they refer to the date of permit. Due to this, it is not certain whether the building has been constructed and also, whether it has been occupied. Another weakness is the fact that the number of permits issued is not necessarily directly related to population change, especially in the case of a decreasing population.

Thus, the use of building permit data also seems to be limited in estimating population for the different geographic areas.

4.3 Unemployment Insurance Commission

The Unemployment Insurance Commission keeps a list of the beneficiaries of the program. A 10% sample of this file has been developed to produce statistics and it could provide migration information. However, this file could hardly be used to estimate migration in Canada. First, a 10% sample of unemployed corresponds to less than 1% of the population. From such a small subpopulation, flows of migrants between provinces could not be derived. Also the non-representativity of that sample (young adults representing a good part of non-employed) calls for suspicion concerning the migration data from that file.

The Commission also maintains a file of wage earners who are contributing to the Unemployment Insurance program. However, no in depth analysis of this file has been done.

4.4 Labour Force Survey

In 1982, Statistics Canada conducted a sample survey of 56,000 households in Canada. The civilian non-institutionalized population aged 15 and over, included in the sample, residing in all provinces were asked a question on their migration history of the past 5 or 6 years. Other valuable information is also available. However, its very small sample (approx. ½% of the population) and the fact that the survey was conducted only once eliminates the Labour Force Survey as migration estimates source.

4.5 Voters' List

Data on voters are generally available in Canada. Federal and provincial election lists could easily be obtained while obtaining municipal lists would necessitate more work. Those lists give information on the number of canadian citizens aged 18+ (landed immigrants are included at the municipal level only). They cover an average 90-95% of the target population. The main shortcoming of that source is that it is not available at regular intervals. Federal and provincial lists are made for elections about every 4 years at dates that are not useful for estimation purposes. It thus seems pointless to consider voters lists.

4.6 Retail Sales

Data on retail sales are collected by Statistics Canada on the basis of sales figures from large stores and from a sample of smaller businesses. These data are collected on a monthly basis and they are made available 3 months after the reference date. These data could be used as a symptomatic indicator of the population change. However, the utility of this data

set seems to be limited in the case of population and migration estimations. This could be due to the fact that retail sales are heavily affected by the economic fluctuations which may not accurately reflect changes in the size of population.

4.7 Trucking Statistics (Moving Companies)

Statistics on a sample of five major moving companies are available in Canada. They cover about 90% of all moves. The interprovincial migration flow could be assessed by weighing the number of reported moves between two different provinces/territories. However, trucking statistics are seriously affected by a time-lag of two years or more.

5. CONCLUDING REMARKS

In this report, an overview of strengths and weaknesses of twelve administrative data files has been presented in order to make recommendations for selecting an alternative data source to the family allowance files. It has been found that there is no file with strengths and weaknesses strictly comparable to those of the family allowance files. However, if the family allowance files cease to be universal, one could suggest the following recommendations:

- Continue further developments in the use of the provincial health insurance file and the Old Age Security records of the federal government in order to produce the total population and migration estimates on a quarterly basis;
- Examine the quality of annual population estimates for the provinces and territories, produced by the Component Method II using the migration estimates from the provincial school enrollment data files; and
- Test the accuracy of the provincial administrative data files (health insurance files, hydro connections, telephone companies and driver's licence) as symptomatic indicators of the population change and the residual net migrants for sub-provincial areas (census divisions and census metropolitan areas in Canada).

REFERENCES

ALMOND, M.M. (1982). An inventory of sources of Canadian migration data. Working Paper, Demography Division, Statistics Canada.

McRAE, D.G. (1985). Use of hydro accounts in the regression population estimates model in British Columbia. Presented at the Federal-Provincial Committee on Demography, Ottawa, Canada.

NORRIS, D.A., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Technical Report, Administrative Data Development Division, Statistics Canada.

NORRIS, D.A. (1983). New sources of Canadian small area migration data. *Review of Public Data Use*, 11-25.

STATISTICS CANADA (Quarterly). *Estimates of Population for Canada, Provinces and Territories.* Catalogue 91-001, Ottawa: Minister of Supply and Services Canada.

STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada, Provinces and Territories.* Catalogue 91-210, Ottawa: Minister of Supply and Services Canada.

STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Regression Method).* Catalogue 91-211, Ottawa: Minister of Supply and Services Canada.

STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Component Method).* Catalogue 91-212, Ottawa: Minister of Supply and Services Canada.

STOCK, R. (1981). *Migration Estimates from Current Administrative Files: Data Sources and Methodologies.* Canadian Plains Research Center, University of Regina.

U.S. BUREAU OF THE CENSUS (1973). *The Methods and Materials of Demography.* Washington, D.C.: U.S. Government Printing Office.

VERMA, R.B.P., BASAVARAJAPPA, K.G., BENDER, R.K. (1984). The regression estimates of population for sub-provincial areas in Canada. *Survey Methodology*, 9, 219-240.

VERMA, R.B.P., BASAVARAJAPPA, K.G. (1985). Recent developments in the estimation of population for small areas in Canada by regression. Presented at the International Symposium on Small Area Statistics, Ottawa, Canada.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 181–186
Statistics Canada

Use of Administrative Data Files for Migration Estimates: A Case Study of Driver's Licence File in Ontario[1]

RAGHUBAR D. SHARMA and CHEUK WONG[2]

ABSTRACT

In Canada, provincial and federal demographers have attempted to use various sets of administrative data to estimate migration flows. This paper presents the development of intra-provincial migration estimates using driver's licence data in Ontario. An evaluation of these migration estimates has been carried out by comparing with those derived from the income tax data by Statistics Canada. Both files provide equally good and complimentary estimates of intra-provincial migration.

KEY WORDS: Administrative files; Population estimates; Component method; Small areas; Error of closure; Intraprovincial migration.

1. INTRODUCTION

Migration is an important component of population projections, and population estimates. As no records regarding the movement of population are kept in Canada, demographers in the federal and provincial governments have attempted to use various sets of administration data to estimate migration flows. Statistics Canada uses revenue data (Norris and Standish 1983), British Columbia utilizes hydro-hookups (McRae 1985), and Alberta uses health care records (Alberta Bureau Statistics 1985). Since 1979, Ontario has been using drivers' licence address changes to estimate intra-provincial migration. Apart from the quality aspect, one major attractiveness of the driver licence data is in its timeliness. There is only a 4 to 5 week time lapse between receiving the data and the date of reference compared with over one and one-half years in revenue data. In this paper we shall present an evaluation of estimates of intra-provincial migration derived from the driver's licence data in Ontario. In the U.S.A., the State of California also uses driver's licence address changes for the estimation of intra-provincial migration (Hoag 1984).

2. DRIVER'S LICENCE DATA FILE

Information on driver's licence address changes is made available by the Ontario Ministry of Transportation and Communications (MTC). A driver is required to notify the Ontario Ministry of Transportation and Communications within 90 days of his/her change of address. The information is available at the postal code area level. These postal code areas can be converted into such subprovincial areas as, counties, regions and municipalities. As Table 1 indicates, data are available for the past seven years. Since 1979, data are also available for each quarter of these years.

More than a million changes of addresses are recorded every year. The majority of these moves tend to be within census divisions (that is, county or regional municipality). However, net inter-county movers averaged only about 22,000 per year. Table 1 indicates that about one-third of the records do not provide a postal code for either origin and/or destination of the mover.

---

[1] Abridged version of a paper presented at the meetings of The Federal-Provincial Committee on Demography, November 28-29, 1985, Statistics Canada, Ottawa.

[2] Raghubar D. Sharma and Cheuk Wong, Sectoral and Regional Policy Branch, Ontario Ministry of Treasury and Economics, Queen's Park, Toronto, Ontario M7A 1Y9.

In Ontario, a person becomes eligible to hold a driver's licence at the age of 16 years. More than 75 per cent of the eligible population holds a driver's licence. The elderly population and female population have a much lower tendency to hold a drivers' licence (Table 2).

3. CONVERSION OF DRIVERS TO MIGRANTS

An adjustment factor is applied to the number of drivers to arrive at the number of movers. This adjustment factor (F) is calculated as follows:

$$FA = \frac{\text{Known and Unknown Movements}}{\text{Known Movements}}$$

$$FB = \frac{\text{Total Population}}{\text{Population with a Licence}}$$

$$F = FA \times FB.$$

Table 1

Number of Total Movers and Number of Movers with Unstated Origin and/or Destination, Ontario, 1975-1985

Year	No. of Known Movers (Inter & Intra Country)	No. of Origin and/or Destination Unstated	Total	% Unstated
1979 (Calendar Year)	881,000	0	881,000	0
1979/80	586,000	301,000	887,000	34
1980/81	566,000	306,000	872,000	35
1981/82	617,000	270,000	887,000	30
1982/83	648,000	259,000	907,000	29
1983/84	822,000	320,000	1,142,000	28
1984/85	831,000	330,000	1,161,000	28

Source: Ontario Ministry of Transportation and Communications.

Table 2

Percent of Population Holding Driver's Licence, Ontario, 1981

	% of Population Holding A Driver's Licence, 1981		
Age	Male	Female	Total
16-19	63.9	36.5	49.1
20-24	92.7	73.0	85.7
25-34	98.6	81.6	90.0
35-44	99.7	79.9	90.0
45-54	96.7	67.8	82.4
55-64	93.2	56.7	74.2
65+	73.1	27.4	46.4
Total	90.6	62.2	75.8

Source: Ontario Ministry of Transport and Communications.

FA accounts for the unstated origins and/or destinations and *FB* accounts for non-driver's licence holders. The factor assumes that migration patterns of those who do not hold driver's licence do not differ from those who hold driver's licence. Similarly, it assumes that migration patterns of those with unstated movements do not differ from those whose movements are stated.

4. INTRA-PROVINCIAL MIGRATION ESTIMATES: DRIVER'S LICENCE VERSUS INCOME TAX FILES

Statistics Canada uses change of address as provided by a taxpayer on his annual income tax return. The number of children are estimated from the number of dependents claimed by the taxpayer. Like the driver's licence data, adjustment factors have to be introduced to the revenue data to overcome unstated postal code and people who do not file an income tax. Furthermore some taxpayers use a non-residential mailing address in their return.

The relative accuracy of migration estimates derived from the income tax file and driver's licence file needs to be tested. Three measures have been applied to test this relative accuracy of the two data sets. They are:

A. Errors of Closure
B. Growth Rates Test
C. Index of Dissimilarity

Ideally, errors of closure and growth rates should be calculated from the population estimates from one census year to the next. Reliable data on driver's licence address changes are available only from 1979 onwards in Ontario. Therefore, 1979 intercensal population estimates of Statistics Canada and estimated 1981 population were used as base. Two sets of population estimates were calculated. First, using the driver's licence address file for intra-provincial migration and second, using the income tax file for intra-provincial migration. All other components, i.e., births, deaths, interprovincial migration and international migration were kept the same for both sets of population estimates.

4.1 Errors of Closure

Two sets of population estimates for the census divisions (one using driver's licence data and the second, using income tax data for intra-provincial migration) were compared with the 1981 census population. The percent difference in the estimated population from the census population is called *error of closure*. Out of 49 census divisions, 23 have smaller errors if driver's licence data are used and 26 census divisions have smaller errors if income tax data are used to estimate intra-provincial migration.

4.2 Toronto Urban Complex

A quite interesting picture emerges in the Toronto Urban complex which includes six regional municipalities (Table 3). Driver's licence data yields a smaller error of closure for the complex as a whole and under-estimates the population for the areas outside of Metro Toronto.

The income tax file gives lower errors of closure for individual census divisions within the complex whereas, for the complex as a whole the error is larger than the driver's licence file (Table 3). Accordingly, driver licence data were used for estimating intra-provincial migration for the Toronto complex as a whole and the distribution to individual census divisions of the complex was based on revenue data.

4.3 Population Growth Rates

Percent change in the population estimates from 1979 to 1981 was calculated for the estimates derived by using the driver's licence file and the income tax file respectively. These growth rates were compared with the 1981 census growth rates.

Table 3
Errors of Closure for Toronto Urban Complex

Census Division	Errors of Closure	
	Income Tax	Driver's Licence
Durham R.M.	−0.21	−0.41
Halton R.M.	0.45	−0.51
Hamilton-Wentworth R.M.	0.10	−0.11
Peel R.M.	−0.63	−1.94
Toronto R.M.	0.01	1.26
York R.M.	−0.11	−5.70
Total Toronto Urban Complex	−0.05	−0.01

There is not much difference in the relative closeness of growth rates of the two sets to census growth rates. The number of census divisions which yield different direction of population change than the census growth rate are 3 for the driver's licence data and 10 for the income tax file (Table 4). This is one aspect where the driver's licence data appeared to yield more reliable estimates than the revenue data.

4.4 Index of Dissimilarity

Index of dissimilarity was calculated for in- and out-migration separately, as the direction of net migration was not the same for some counties for the two sets of estimates. The value of the index of dissimilarity can vary between 0 and 100. It is the half of the sum of the absolute differences between the two corresponding percent distributions and is equivalent to the sum of the positive differences or the sum of the negative differences (Shryock and Siegel 1971). The general formula is:

$$ID = \tfrac{1}{2} \sum |r_2 - r_1|$$

where, r_2 and r_1 are the corresponding percentages in the two distributions.

The low values of the index indicate that both files (the driver's licence and the income tax) yield quite similar estimates of intra-provincial migration for the census divisions of Ontario. However, over the four years the extent of dissimilarity increases for out-migration and improves for in-migration (Table 5).

5. CONCLUSION AND SUMMARY

This study attempts to compare the intra-provincial migration estimates derived from the driver's licence file with those derived from the income tax file. Both files provide reasonably good measures of the magnitude of intra-provincial migration for the Census Divisions of Ontario.

Although the driver's licence data appeared to provide better estimates in the direction of intra-provincial migration, the income tax data resulted in slightly more counties with smaller errors of closure and in addition yielded somewhat better results in some major areas (for example, distribution within the Toronto/Hamilton urban complex). In view of their respective strengths, the appropriate approach is to combine the use of these two data sources.

Another issue that should be noted is that the evaluation was based on three years only i.e., 1979 to 1981. A more accurate assessment on the quality of these two data files cannot be made until the availability of the 1986 census data.

To further improve the quality and the applications of the driver's licence data, the following two areas are suggested for further research:

- Verification of *FA* factor based on actual counts of unknown origin/destination through using manual coding of addresses.
- Extension of the use of the driver's licence data file as an additional source to family allowance and revenue data for inter-provincial migration estimates.

The driver's licence file tends to over-estimate migrants for Metro Toronto and underestimate for the areas surrounding Metro Toronto. The reverse seems true for the income tax file. For this region as a whole, the driver's licence file gives better estimates for intra-provincial migration than the income tax file. The income tax file provides a better distribution of intra-provincial migrants in the counties of this region.

Table 4

Census Divisions Which Yield Different Direction of Population Change Than Those Based on Census

Census Division	% Changed Based On	
	Income Tax	Census
Bruce	0.11	−1.77
Grey	−0.04	0.17
Hastings	0.12	−1.03
Leeds and Grenville	0.21	−0.32
Niagara	0.15	−0.46
Northumberland	0.31	−0.82
Oxford	0.17	−0.09
Parry Sound	1.20	−0.51
Stormond/Dundas/Glengarry	0.44	−0.17
Sudbury T.D.	0.41	−2.28
Province	1.60	1.46
	% Change Based On	
	Driver's Licence	Census
Leeds and Grenville	0.02	−0.32
Parry Sound	0.81	−0.51
Thunder Bay	0.42	−0.20
Province	1.60	1.46

Table 5

Index of Dissimilarity

Year	Index of Dissimilarity	
	In-Migration	Out-Migration
1979-80	5.61	3.50
1980-81	5.54	3.82
1981-82	5.26	4.82
1982-83	4.41	4.87

REFERENCES

ALBERTA BUREAU OF STATISTICS (1985). The development of Alberta health care records and their application to small area population estimates. A paper presented at the Meetings of the Federal-Provincial Committee on Demography, Ottawa.

HOAG, ELIZABETH (1984). Estimating annual migration for California counties using driver's licence address change. A paper presented at the Meetings of the Population Association of America, Minneapolis, Minnesota.

McRAE, DONALD G. (1985). The use of hydro accounts in the British Columbia based population estimates. A paper presented at the Meetings of the Federal-Provincial Committee on Demography.

NORRIS, D., and L. STANDISH (1983). A technical report on the development of migration data from taxation records. Administrative Data Development Division, Statistics Canada.

SHRYOCK, HENRY M., and SIEGEL, JACOB S. (1971). *The Methods and Materials of Demography*. Washington: U.S. Bureau of Census.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 187–196
Statistics Canada

The Development of Alberta Health Care Records and Their Application to Small-Area Population Estimates[1]

F. AHMAD, R. CHOW, O. DEVRIES,
A. HASHMI, and M. MARCOGLIESE[2]

ABSTRACT

This paper examines the use of administrative files from Alberta's Health Care Insurance Plans combined with Vital Statistics data as inputs for estimating population. Results, which are presented and compared with Census data, indicate that Health Care data can be used to produce accurate population estimates at the provincial level and for smaller areas such as census divisions and municipalities.

KEY WORDS: Administrative files; Component method; Small areas; Residual net migration.

1. BACKGROUND

During the mid to late 1970's, the Province of Alberta experienced rapid economic growth led by activity in the oil and gas industry, which generated high population growth. Governments, in order to effectively provide goods and services for the influx of people into various regions, required timely data on where and by how much population was growing. With the need for up-to-date population data, it was felt that the federal quinquennial census was not sufficiently frequent nor current (census data are released about twelve to eighteen months after the reference year). Consequently, provincial agencies, and in particular, the Alberta Bureau of Statistics, began investigating alternative sources of timely population data.

After examining a number of potential sources, the Bureau began assessing administrative health care insurance data from the Alberta Health Care Insurance Plan (AHCIP) files to develop population statistics. The remainder of this paper highlights work undertaken by the Bureau to develop the AHCIP records and to use the data in estimating small-area population.

2. DEVELOPMENT OF AHCIP RECORDS INTO HEALTH CARE COUNTS

This section describes briefly the nature of the AHCIP records and evaluates the counts developed.

2.1 Developing Health Care Counts Data

The Bureau receives selected registration records via computer tape, on a quarterly basis, from the AHCIP registration-billing system. (The tape contains only a partial listing, in particular, all names, identifiers, etc. have been stripped such that the confidentiality of all individuals is strictly preserved.) The file contains information such as addresses, postal codes, registration and cancellation dates, age and sex for every registrant. (A detailed description of the record layout is available upon request.)

---

[1] Abridged version of the paper presented at the Federal-Provincial Committee on Demography meeting held on November 26-27, 1985, Ottawa, Canada.

[2] F. Ahmad, R. Chow, O. DeVries, A. Hashmi and M. Marcogliese, Alberta Bureau of Statistics, Alberta Treasury, Sir Frederik W. Haultain Building, 9811-109th Street, Edmonton, Alberta, Canada T5K 0C8.

The reporting unit of the AHCIP file is the registration. Each registration may contain up to twenty-five individuals; one registrant (usually the person who pays the premiums) and up to twenty-four dependents. There are currently about 1.7 million active registrations accounting for roughly 2.6 million individuals. In addition, the file is historical and includes all individuals ever covered under AHCIP since its inception in 1969.

The file is processed through four phases.

a) Edit–notes and/or corrects errors according to edit check criteria.

b) Purge–uses the edited raw data file and selects active individuals.

c) Consolidation–matches postal codes between the purged file and the Bureau's Postal Code Translator File (PCTF) and attaches the geographic reference information to the AHCIP records.

d) Aggregation–takes the consolidated file and aggregates males and females by single years of age for each postal code. This reduces the number of records/individuals from approximately 2.6 million to fewer than 120,000 and significantly reduces the subsequent systems processing costs.

The aggregated file is used for the production of age and sex counts by any geographic area definable through the 60,000 PCTF Alberta codes.

2.2 Evaluation of the Counts Data

To evaluate the health care counts data, Census of Canada population figures for 1976 and 1981 were used for comparison. The 1981 AHCIP records were considered to be more accurate than the 1976 file, therefore, the evaluation relied more heavily upon the 1981 census comparisons. Also used as a second basis of comparison were municipal censuses data, even though these data generally were not considered to be as reliable as Canada Census figures. The municipal censuses, however, provided insight into the magnitude of the variations as well as the relative distributions of age, sex and trends (growth or decline) over time. An additional source of comparison was intercensal population estimates prepared by the Bureau and by Statistics Canada.

Basic findings:

a) On a provincial basis, AHCIP counts overestimate both Canada Census and total municipal censuses figures by about 3.5% to 4.5%. Age and sex distributions are more accurate and the correlation coefficients indicate consistency of trends (over/under estimates) over time.

b) At the census division (CD) level, AHCIP counts varied from Canada Census figures from −2.6% to 9.7% (see Table 1). Comparisons with intercensal population estimates indicated a similar variance. As with the provincial level data, age and sex distributions and the trend consistency proved highly reliable.

c) At the census consolidated subdivision (CCSD) level, for fifty of the seventy-one CCSDs, health care data were within ± 10% of the Census counts. The largest discrepancy was −56.5% (Municipal District 135).

Most problem areas had major urban centres located close to the county, municipal district and improvement district boundaries. No specific anomalies were found when testing the age and sex distributions, although relationships were not as strong as with the province and the census division levels.

d) At the census subdivision (CSD) level, preliminary figures showed discrepancies between the AHCIP counts and 1981 Census data ranged from −100% to +955%.

Consequently, the twenty-eight largest areas of over 5,000 in population were used at the CSD level. The six largest CSDs (Edmonton, Calgary, Lethbridge, Medicine Hat, Red Deer and St. Albert) displayed overcounts ranging from 3% to 9%. Eight other CSDs differed up to ±20%, while sixteen showed somewhat greater than ±20% variation. Again, no specific age and sex distribution anomalies were detected, although discrepancies were greater than those at more aggregated levels. As well, twenty-seven of the twenty-eight CSDs indicated high trend consistency.

As the geographic area decreases in size, AHCIP counts become less reliable; age and sex distributions, although less accurate, still remain strong; and trend consistency (counts over time) remain highly correlated with a few notable exceptions. The limitations of AHCIP counts as population indicators primarily can be attributed to one of two main sources: a) the AHCIP administrative procedures/inaccuracies; or b) use of postal codes.

a) *AHCIP Administrative Procedures:*

1) As an insurance programme, a chief concern is to supply coverage. Therefore, efforts are directed to getting people onto the system to ensure universal coverage with less effort placed on getting individuals off the system. This has resulted in more people being registered than are actually in the province.

Table 1

Comparisons of Alberta Health Care Counts and Canada Census Data for Alberta Census Divisions

	Year							
	1976				1981			
Census Division	Census Count	AHCIP Count	Percent Difference Count	Actual Difference Count	Census Count	AHCIP count	Percent Difference Count	Actual Difference Count
1	46,990	45,789	−2.56	−1,201	55,375	55,748	0.67	373
2	96,995	97,229	0.24	234	110,477	111,567	0.99	1,090
3	32,898	33,884	3.00	986	35,652	36,463	2.27	811
4	12,130	12,101	−0.24	−29	12,119	12,038	−0.67	−81
5	35,424	35,656	0.65	232	38,382	38,457	0.20	75
6	524,554	538,432	2.65	13,878	668,682	699,999	4.68	31,317
7	37,866	38,235	0.97	369	40,071	40,359	0.72	288
8	95,384	95,063	−0.34	−321	123,642	124,666	0.83	1,024
9	19,903	21,832	9.69	1,929	21,670	23,338	7.70	1,668
10	67,171	67,168	0.00	−3	78,417	78,532	0.15	115
11	632,909	646,799	2.19	13,890	762,041	796,884	4.57	34,843
12	63,129	62,011	−1.77	−1,118	84,221	86,183	2.33	1,962
13	46,305	47,258	2.06	953	53,701	54,282	1.08	581
14	19,386	21,039	8.53	1,653	24,635	25,991	5.50	1,356
15	106,993	111,678	4.38	4,685	128,639	134,451	4.52	5,812
Unknown[a]		48,462				19,279		
Alberta	1,838,037	1,922,636	4.60	84,599	2,237,724	2,338,237	4.49	100,513

[a] Unknown, represent counts without address identifiers.

Source: Statistics Canada 1976 and 1981 Censuses; Alberta Health Care Insurance Plan data, prepared by Alberta Bureau of Statistics, Alberta Treasury.

2) Mailing addresses are used rather than residential addresses, which has created difficulties in assigning geographic locations. Discrepancies occur in areas where significant rural populations surround an urban centre and the rural populace pick up their mail in the urban centre. Consequently, most urban areas are overcounted while rural areas are undercounted.
3) Incomplete and inaccurate data, especially related to postal codes, make it difficult to produce small-area statistics due to undercounting.
4) Time lags in reporting and recording of the data influence counts. Generally speaking, it takes three to six months to get an individual onto the system (birth, in-migrant) but it requires usually much longer to be removed from the active system (death, out-migrant). The lags, however, are difficult to follow and differ substantially depending on the circumstances.

b) *Postal Codes:*

1) Postal codes define delivery service areas (where a person gets his mail), not necessarily a residence. This factor limits the accuracy of assigning AHCIP registrations to appropriate geographic areas. In particular, it creates urban-rural split problems, as discussed.
2) A six-digit postal code, by itself, is not always enough to determine the service delivery area. A rural route, suburban service, or box number may be required to further specify a more exact location.
3) Postal codes have been insufficient, especially in rural areas, to aggregate to appropriate levels. For example, there are approximately 363 census subdivisions in Alberta, but the Bureau's PCTF can derive only 324 of these.

The problems outlined above have precluded the release of AHCIP counts as approximations of actual population. Although the counts were quite good in some areas, in others, they were poor or inconsistent. With the strong relationships between health care, age and sex distributions and those of Canada Census, as well as the consistency of trends over time, the counts have been used in conjunction with the Bureau's population estimation methodology (as discussed in the next section).

3. APPLICATION OF HEALTH CARE COUNTS TO SMALL-AREA POPULATION ESTIMATES

The Bureau has produced intercensal population estimates for Alberta and provincial census divisions for nearly a decade. During this period, various methodologies and data sources have been examined and used to improve the quality of these estimates. To date, significant success has been achieved with the component method using health care counts as input data. These data have been used to derive the age and sex structure of the Alberta population at the provincial and census division level and to produce provincial and census division population estimates. Also, recently, the data have been used to test the applicability in preparing census subdivision population estimates.

3.1 Estimation Methodology

The estimation methodology employed by the Bureau to produce subprovincial population estimates is comprised of two parts. Part one presents the method of estimating migrant population. Part two outlines the method used to develop population estimates.

a) *Estimating Migrant Population Using Health Care Counts*

The Bureau developed data from three administrative files: counts from AHCIP records; births from data supplied by Alberta Vital Statistics; and deaths, also supplied by Alberta Vital Statistics. These sources were used to calculate net migration. Basically for any small area, the growth of health care counts is obtained from the differences in counts between time *t* and time *t-1*. This residual less the area's natural increase (births minus deaths) calculates the inflow (or outflow) of individuals, i.e., net migration. This procedure is mathematically expressed as:

$$HMIG = [(HC_t - HC_{t-1}) - (B - D)]$$

Where:

$HMIG$ = health care net migration counts between time t and $t-1$

HC_t = total health care counts at time t

HC_{t-1} = total health care counts at time $t-1$

B = total births during time interval t to $t-1$

D = total deaths during time interval t to $t-1$.

This health care migrant population estimate, however, is subject to the same over and under counting difficulties discussed in Section 2. As a result, although this approach would prepare estimates for small areas at the provincial level, these estimates would be less reliable than the provincial migration estimates currently derived using interprovincial flows to family allowance recipients. (The family allowance files are also used by Statistics Canada, which ensures provincial estimates generally are consistent with those produced at the federal level.)

To further improve the small-area migration estimates and to ensure consistency with estimates at the provincial level, should the small areas be aggregated to a provincial total, an adjustment using a ratio distribution was encompassed. With this approach, the ratio of net migration from health care counts for an area over the net migration from health care counts for the province is multiplied by the provincial net migration calculated in connection with the Bureau's quarterly population estimates. Mathematically, the equation is:

$$AMIC_i = \frac{HMIG_i}{HMIG_a} \times PMIG$$

Where:

$AMIG_i$ = adjusted net migration in area i

$HMIG_i$ = health care net migration of counts for area i

$HMIG_a$ = health care net migration of counts for Alberta

$PMIG$ = estimated provincial net migration from Alberta's quarterly population estimates.

This adjusted migration estimate (AMIG) is then used as input into estimating population.

b) *Estimation of Population For Small Areas*

The adjusted estimated net migration (AMIG) for each area is used in an equation using the components of population growth (births, deaths and migration):

$$P_{i_t} = P_{i_{t-1}} + (B_i - D_i) + AMIG_i$$

Where:

P_{i_t} = estimated population in area i at time t

$P_{i_{t-1}}$ = population in area i at time $t-1$

3.2 Evaluation of Small-Area Estimates

Using the above approach, the Bureau has developed population estimates for Alberta's fifteen census divisions and twenty-eight municipalities with populations over 5000. The results, so far, have been promising.

The results of a comparison between 1981 census data and estimates for 1981 prepared with 1976 census figures as a base population using the above described methodology, at the census division level, are presented in Table 2. For thirteen of the fifteen divisions the estimates were within ±2.0% variation compared to the 1981 census. Only the two smallest CDs (9 and 14) showed a five-year deviation greater than 2.0%. The average absolute deviations (i.e., average annual deviations) were no greater than 0.5% for all census divisions.

The twenty-eight population estimates for municipalities were compared to the 1981 census counts, as well as available data from municipal censuses conducted from 1982 to 1984 (Tables 3 and 4). Federal census comparisons showed nineteen estimates of the twenty-eight municipalities had an average absolute deviation of less than ±1.0%. Only six municipalities had annual differences greater than 2.0%. Comparisons with municipal censuses conducted between 1982 and 1984, yielded twenty-two instances of deviations within ±1.0%, fourteen ranging between ±1.0% and ±3.0%, while nine had deviations greater than ±3.0%.

In general, the estimation results have been satisfactory and encouraging. The development of AHCIP registrant counts and the component approach employed to estimate population have improved the accuracy of the population estimates produced and opened up possibilities for deriving estimates for user-defined small geographic areas. The Bureau will continue to investigate ways to improve the AHCIP counts (some of which are related to new administrative procedures being incorporated for the AHCIP). Also, the population estimation methodology will be further refined as new data techniques become available.

3.3 Summary of Advantages and Disadvantages of Using AHCIP

Using health care counts in deriving small-area population estimates has a number of advantages and disadvantages.

Table 2

Comparisons of Canada Census Counts and Alberta Bureau of Statistics Population Estimates for Alberta Census Divisions

Census Division	Census 1976	Bureau Estimates[a]			
		Natural Increase[b] 1976-81	Net Migration 1976-81	Growth 1976-81	Population 1981
1	47,000	2,730	6,080	8,810	55,810
2	96,980	6,120	7,190	13,310	110,290
3	32,870	2,310	100	2,410	35,280
4	12,140	490	−520	−30	12,110
5	35,460	1,820	790	2,610	38,070
6	524,570	33,860	107,540	141,400	665,970
7	37,820	2,010	−10	2,000	39,820
8	95,400	6,140	20,860	27,000	122,400
9	19,850	1,040	200	1,240	21,090
10	67,230	1,650	8,550	10,200	77,430
11	632,830	43,880	90,880	134,760	767,590
12	63,130	6,470	16,130	22,600	85,730
13	46,300	2,040	4,320	6,360	52,660
14	19,450	2,200	2,430	4,630	24,080
15	107,010	10,260	10,040	20,300	127,310
Alberta	1,838,040	123,020	274,580	397,600	2,235,640

Census Division	Census 1981	Difference		Average Absolute Deviation
		Number	%	
1	55,360	450	0.81	0.16
2	110,470	−180	−0.16	0.03
3	35,640	−360	−1.01	0.20
4	12,120	−10	−0.08	0.02
5	38,430	−360	−0.94	0.19
6	668,680	−2,710	−0.41	0.08
7	40,030	−210	−0.52	0.10
8	123,690	−1,290	−1.04	0.21
9	21,630	−540	−2.50	0.50
10	78,390	−960	−1.22	0.24
11	762,080	5,510	0.72	0.14
12	84,220	1,510	1.79	0.36
13	53,690	−1,030	−1.92	0.38
14	24,650	−570	−2.31	0.46
15	128,640	−1,330	−1.03	0.21
Alberta	2,237,720	−2,080	−0.09	0.02

[a] Data are experimental.
[b] Natural increase refers to the number of births minus the number of deaths.
Note: Components may not add to total due to rounding.
Source: Statistics Canada 1976 and 1981 Censuses; Alberta Bureau of Statistics Estimates.

Table 3

Comparisons of the Canada Census Counts and Alberta Bureau of Statistics Population Estimates for Selected Alberta Municipalities

Municipality	Census 1976	Bureau Estimates[a] Natural Increase[b] 1976-1981	Net Migration 1976-1981	Population 1981	Census 1981	Difference %	Average Absolute Deviation
Airdrie	1,410	580	5,090	7,070	8,410	−15.9	3.2
Brooks	6,340	730	2,370	9,440	9,420	0.2	0.0
Calgary	469,920	30,310	93,760	593,990	592,740	0.2	0.0
Camrose	10,100	150	2,570	12,830	12,570	2.1	0.4
Crowsnest Pass	5,250	40	−410	4,880	7,310	−33.2	6.6
Drayton Valley	4,300	530	1,760	6,590	5,040	30.8	6.2
Drumheller	6,150	20	220	6,390	6,510	−1.8	0.4
Edmonton	461,360	27,900	51,240	540,510	532,250	1.6	0.3
Edson	4,040	510	2,490	7,040	5,840	20.5	4.1
Fort McMurray	15,420	2,900	14,140	32,460	31,000	4.7	0.9
Fort Saskatchewan	8,300	800	2,660	11,760	12,170	−3.4	0.7
Grande Prairie	17,630	1,970	6,300	25,900	24,260	6.8	1.4
Hinton	6,730	760	−820	6,670	8,340	−20.0	4.0
Innisfail	2,900	230	1,930	5,060	5,250	−3.6	0.7
Lacombe	3,890	150	1,210	5,240	5,590	−6.3	1.3
Leduc	8,580	920	3,430	12,930	12,470	3.7	0.7
Lethbridge	46,750	2,070	4,400	53,220	54,070	−1.6	0.3
Medicine Hat	32,810	1,770	6,010	40,590	40,380	0.5	0.1
Peace River	4,840	580	970	6,390	5,910	8.1	1.6
Ponoka	4,640	−10	530	5,160	5,220	−1.1	0.2
Red Deer	32,180	2,300	11,790	46,270	46,390	−0.3	0.1
Spruce Grove	6,910	1,110	4,710	12,730	10,330	23.2	4.6
St. Albert	24,130	2,360	6,670	33,160	32,000	3.6	0.7
Stettler	4,180	500	580	5,270	5,140	2.5	0.5
Taber	5,300	320	410	6,020	5,990	0.5	0.1
Vegreville	4,160	80	860	5,090	5,250	−3.0	0.6
Wetaskiwin	6,750	300	2,440	9,490	9,600	−1.1	0.2
Whitehorse	3,880	600	1,150	5,630	5,590	0.7	0.1
Alberta	1,838,040	123,020	274,580	2,235,630	2,237,720	−0.1	0.0

[a] Data are experimental.
[b] Natural increase refers to the number of births minus the number of deaths.
Note: Components may not add to total due to rounding.
Source: Statistics Canada 1976 and 1981 Censuses; Alberta Bureau of Statistics Estimates.

Table 4

Comparisons of Alberta Municipal Censuses and Alberta Bureau of Statistics Population Estimates for Selected Municipalities

Municipality	1982			1983			1984		
	Bureau Estimate[a]	Municipal Census	Deviation %	Bureau Estimate[a]	Municipal Census	Deviation %	Bureau Estimate[a]	Municipal Census	Deviation %
Airdrie	9,450	9,980	−5.3	9,830	10,430	−5.8	10,080	--	--
Brooks	9,640	--	--	9,790	--	--	9,510	--	--
Calgary	614,930	623,130	−1.3	622,510	620,690	0.3	615,140	619,810	−0.8
Camrose	12,880	12,810	0.6	12,970	--	--	13,070	12,750	2.5
Crowsnest Pass	7,490	7,580	−1.1	7,530	--	--	7,350	--	--
Drayton Valley	5,120	4,870	5.2	5,200	--	--	5,310	4,920	7.9
Drumheller	6,660	--	--	6,700	6,670	0.4	6,620	--	--
Edmonton[b]	550,930	551,310	−0.1	557,400	560,090	−0.5	551,140	--	--
Edson[b]	6,110	6,290	−2.9	6,220	--	--	6,080	7,110	−14.5
Fort McMurray	32,930	33,580	−1.9	33,600	34,490	−2.6	35,150	35,350	−0.6
Fort Saskatchewan	12,530	12,460	0.6	12,650	12,470	1.4	12,620	--	--
Grande Prairie	24,650	--	--	24,910	24,080	3.5	25,370	24,410	3.9
Hinton	8,820	8,820	0.0	8,980	8,830	1.8	8,950	8,900	0.6
Innisfail	5,420	5,440	−0.4	5,460	--	--	5,440	5,440	0.0
Lacombe	5,810	5,720	1.5	5,850	5,850	5,950	−1.8	5,850	--
Leduc	12,880	--	--	13,010	--	--	13,290	--	--
Lethbridge[b]	55,440	56,500	−1.9	55,900	58,090	−3.8	57,500	--	--
Medicine Hat[b]	41,070	--	--	41,440	42,270	0.7	41,540	--	--
Peace River	6,080	--	--	6,150	--	--	6,250	--	--
Ponoka	5,310	--	--	5,310	--	--	5,280	--	--
Red Deer	48,450	48,560	−0.2	49,230	50,260	−2.0	50,860	51,070	−0.4
Spruce Grove	11,080	10,780	2.7	11,410	11,310	0.9	11,550	11,570	−0.1
St. Albert	33,170	32,980	0.6	33,740	35,030	−3.7	34,840	35,530	−1.9
Stettler	5,180	--	--	5,220	--	--	5,300	--	--
Taber	6,140	--	--	6,210	--	--	6,360	6,380	−0.4
Vegreville	5,280	5,250	0.6	5,290	--	--	5,390	--	--
Wetaskiwin	9,880	9,900	−0.2	9,990	10,020	−0.3	10,080	--	--
Whitecourt	5,710	--	--	5,840	--	--	5,710	--	--

[a] Data are experimental.
[b] Annexation took place between 1982 and 1984.
Note: "--" indicates that a municipal census is not available.
Source: Alberta Municipal Affairs, 1982-1984 Municipal Censuses; Alberta Bureau of Statistics Estimates.

Advantages:

a) AHCIP registration data provides universal coverage of all individuals in Alberta;
b) Registration lag appears to be random and does not adversely affect distributions or trends of the counts;
c) Data are available on a timely/frequent basis; and
d) The file contains some socio-economic information on registrants and dependents (e.g., age, sex and marital status) to enable the production of more than basic population estimates.

Disadvantages:

a) Residency based on postal codes can lead to some inaccuracies;
b) AHCIP registrants can leave the system, for example, death and out-migration, without notifying AHMC resulting in overcounts; and
c) Administrative procedures may cause discrepancies/inaccuracies in the number of Alberta Health Care registrants.

4. CONCLUSION

Our experience with health care development has been very positive. The greatest potential is the use of the counts in a component model to produce estimates for small areas as well as the excellent age-sex distribution ratios and trend consistency. Costs of development of the demographic reporting systems were not considered excessive in light of these benefits. For other provincial agencies contemplating the development of provincial health care files, the Bureau would certainly be willing to discuss its experiences in more detail and make available additional information, such as record layouts and system processing costs.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 197–201
Statistics Canada

The Use of Hydro Accounts in the British Columbia Regression Based Population Estimation Model[1]

DONALD G. McRAE[2]

ABSTRACT

The accuracy of small area population estimates derived from a regression based model is heavily dependent on the ability of the indicator data selected to accurately reflect population change. Hence, prior knowledge as to the characteristics of the administrative data used as potential population indicators in a regression model is important. This report summarizes the strengths and weaknesses associated with the use of residential hydro accounts in the British Columbia regression based population estimation model.

KEY WORDS: Small area population estimates; Regression method; Difference-correlation method; Population indicators; Hydro accounts; Family allowance recipients.

1. INTRODUCTION

The Central Statistics Bureau produces post-censal population estimates for a variety of geographic units within the Province of British Columbia including municipalites, local health areas, census divisions and RCMP regions among others. Current population estimates are produced for these sub-provincial areas by means of a regression approach, specifically the Difference-Correlation Method (DCM).

A detailed description of this methodology is given in earlier papers (Central Statistics Bureau 1982, McRae 1985). The data used as indicators of population are the number of family allowance recipients (F), and/or the number of residential hydro accounts (H). The characteristics of this second data source, residential hydro accounts, relative to the British Columbia model will be examined over the remainder of this paper.

2. DATA SOURCES

Residential hydro accounts data within British Columbia are obtained from nine different organizations. These are:

Organization	% of Total Hydro Accounts (1985)
(1) British Columbia Hydro	90.9
(2) West Kootenay Power and Light	4.7
(3) Princeton Light and Power Co.	0.2
(4) City of Kelowna	0.8
(5) City of Penticton	0.8
(6) District of Summerland	0.3
(7) City of New Westminster	1.7
(8) City of Grand Forks	0.1
(9) City of Nelson	0.6

[1] Presented at the meeting of the Federal-Provincial Committee on Demography, Ottawa, November 28-29, 1985.

[2] D.G. McRae, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia, 2nd Floor, 1405 Douglas Street, Victoria, British Columbia, Canada V8W 3C1.

The views expressed in this paper are those of the author and do not necessarily represent the views of the Government of British Columbia.

The major suppliers of residential electrical power are British Columbia Hydro and West Kootenay Power and Light. The other organizations purchase power from the two major suppliers, and retail this electricity to their own customers (usually the residents of the municipality).

3. DATA FORMAT

Of the nine sources of residential hydro accounts data, only that provided by British Columbia Hydro is in machine readable form. The other eight organizations provide the data totalled by municipality (urban), along with a total of any rural (non-municipal) customers. The reference date for all data is the May 31 billing file, and in most cases the data can be obtained within 2 to 3 weeks of the billing date.

Data provided by British Columbia Hydro is in two formats. The first shows the number of residential meters as of May 31 by Capital District Code. A Capital District Code, of which there are approximately 248 in the Province, is an administrative unit used by British Columbia Hydro and corresponds to a municipality where municipalities exist. By agreement, both the major power suppliers in the Province pay each municipality a certain percentage of the annual revenue collected from the residential customers in that municipality in lieu of property taxes. As a result, power companies such as British Columbia Hydro, design their accounting systems to correspond to customers within municipal boundaries. In addition, British Columbia Hydro attempts to maintain a close correspondence between Capital District boundaries and school district boundaries.

The second format provides for each of the one million plus residential meters the postal code of billing address. This second data file allows the easy translation of hydro meters to geographic units other than municipalities and school districts via the postal code.

4. STRENGTHS OF THE HYDRO DATA IN A REGRESSION MODEL

Empirical tests of the two different data sources, hydro meters (H) and family allowance accounts (F), were conducted by producing 1981 population estimates with each separately and together. The regression coefficients used were derived from the pooled 1971/76 and 1976/81 periods, and the base year was 1976. The results were compared with the 1981 Census, and the Average Absolute Percent Errors (AAPE) were calculated. The results are given in Tables 1 and 2.

As can be seen in Table 1, population estimates based on hydro data produce, on average, lower percentage errors than the family allowance based estimates. Closer examination of Table 1 reveals that the improvement in estimation accuracy lies almost entirely with the estimates for areas with population less than 4000. This observation is reinforced in Table 2 where it is shown that, statistically speaking, there is a significant difference in the estimation accuracy between the hydro and family allowance based estimates for areas less than 4000 population.

The marginal effect of adding another population indicator to the Difference-Correlation Method can be judged by examining the change in estimation accuracy with and without the additional indicator. It would appear from Tables 1 and 2 that the inclusion of hydro data statistically improves the estimation accuracy in both large and small areas. Family allowance data, on the other hand, improves the accuracy for larger areas but reduces the estimation accuracy for smaller areas, with no statistically significant effect overall.

Table 1

Comparison of Estimation Errors Among Data Sources for British Columbia Municipalities - 1981

Data Source	AAPE Overall	n	Population ≥ 4000 AAPE	n	Population < 4000 AAPE	n
DCM/H/F	5.53	158	2.99	88	8.72	70
DCM/H	5.16	158	4.04	88	6.58	70
DCM/F	10.46	167	4.57	92	17.69	75

$$\text{AAPE} = \left[\sum_{i=1}^{n} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| \right] \div n \times 100$$

where:

Y_i = census population for region i
$\hat{Y}_i$ = estimated population for region i
n = number of areas estimated.

Table 2

Test for Statistically Significant Differences Between the Average Absolute Percent Errors for Selected Data Sources - 1981

Data Source	95% Confidence Interval for the Average Difference in Asbolute Percent Errors: Overall	Population ≥ 4000	Population < 4000
DCM/H/F - DCM/H	.37 ± .86	−1.05 ± .56[a]	2.14 ± 1.76[a]
DCM/H/F - DCM/F[b]	−4.86 ± 1.55[a]	−1.57 ± .85[a]	−9.00 ± 3.11[a]

[a] Statistically significant differences at the 5% level utilizing a two tailed T-test, paired samples and assuming normally distributed means.

[b] In order to pair the samples only 158 of the possible 167 family allowance base estimates were used. The number of observations were: overall, 158; greater than or equal to 4000, 88; less than 4000, 70.

5. WEAKNESSES OF HYDRO DATA IN A REGRESSION MODEL

One problem encountered when using hydro data in a regression model for population estimates is that of vacant dwellings, or more accurately, significant differences in the rate of vacant dwellings between the base and estimating years. This weakness of the data was demonstrated by the 1981 evaluation of the communities in the Peace River-Liard region of British Columbia (McRae 1982). As a result of the North-East Coal project, the communities of the Peace River-Liard Census Division in 1981 were experiencing a building boom as developers contructed dwellings in anticipation of a population influx. Each dwelling, occupied or not (or even under construction) would require a meter, which may have had low usage, but was still active and hence counted. As a result, the change in share of meters from 1976 to 1981 was overstated relative to population, producing overestimates of the 1981 population for many of Peace-River Liard communities.

Another weakness of the hydro data is the potential for a change-over of multiple dwelling units from single to multiple meters. This may occur when an older apartment building, for example, serviced by a single meter is remodelled or replaced with individually metered units. This problem would produce an overestimate of population in a regression model if it were to occur sometime between the base and estimating years.

Finally, some problems will result if the hydro data is used for areas that have a changing nature, or in other words, a changing relationship between population and residential meters. One example of this in B.C. is the resort community of Whistler. Fifteen years ago this municipality was largely a collection of winter cabins on a ski hill. However, over the last decade this area has been shifting to a year round residence basis. Consequently, the number of persons per hydro meter, which was originally very low relative to the B.C. average, is moving toward the norm. Like the vacancy problem, the use of hydro data to estimate the population for such a community would likely result in above average errors.

The solution to all three of the problems mentioned above is to remove accounts that have a low monthly or bi-monthly usage, and hence are assumed to be vacant. The feasibility of this procedure is currently being examined by the Central Statistics Bureau in relation to the data obtained from B.C. Hydro. If possible, we hope to have the improved data set available for calibration against the 1986 Census. Currently, as a partial solution hydro data for areas that had in 1981 a low or high ratio of persons per meter relative to the provincial norm (i.e. less than 2 or greater than 5) are not used.

A final potential weakness of the hydro data is the reliance on external and different organizations for the information. In the past, this situation generally has not proven to be a problem. However, there have been some rare cases that have called into question the quality of the meter data collected in the field. Such a case may be a boundary change of a municipality not being reflected in the meter data, or the addition of some types of non-residential accounts (such as lamp standards) to the data. As a result, careful monitoring of the data is important.

6. CONCLUSIONS

The following strengths and weaknesses are associated with the use of hydro data in the British Columbia regression based population estimation model.

Strengths:

(a) The hydro data, when used in a regression model, produces a lower average absolute percent error than family allowance data for small areas.

(b) The data is obtained from each supplier in a format that is already aggregated to municipalities. The major advantage of this is that changes in municipal boundaries, which occur regularly, are reflected in the data with no additional work on the part of the Bureau.

(c) The majority of the data can be obtained in machine readable form along with the postal code. This allows the easy translation of the data to geographic regions other than municipalities when sorted by the Bureau's postal code Translation Master File.

(d) The data can be obtained free of charge from each of the suppliers within a relatively short time period (2 to 3 weeks).

Weaknesses:

(a) Differential vacancy rates between the base and estimating years will bias the estimates.

(b) Dwelling units (such as apartments) that change from a single to multiple meter sometime between the base and estimating years will bias the estimates upwards.

(c) Areas with a changing nature, such as from a seasonal to "stable" population, will introduce bias into the estimates.

(d) The data is obtained from external and different organizations. This potentially could cause problems in terms of data quality and comparability, as well as producing a situation in which the priorities of the Bureau's population estimates program are subservient to the administrative needs of an external organization.

REFERENCES

CENTRAL STATISTICS BUREAU (1982). British Columbia municipal population estimation methodology. Unpublished Report, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia.

McRAE, D. (1985). A regression approach to small area population estimation. Paper submitted to the International Symposium on Small Area Statistics, Ottawa, Canada.

McRAE, D. (1982). British Columbia small area estimation model - 1981 municipal and census division evaluation. Unpublished Report, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 203–210
Statistics Canada

Estimating the Age/Sex Distribution of Small Area Populations[1]

DAVID S. O'NEIL and CHRIS D. McINTOSH[2]

ABSTRACT

This paper describes a method of producing current age/sex specific population estimates for small areas utilizing as inputs total population estimates, birth and death data and estimates of historical residual net migration. An evaluation based on the 1981 Census counts for census divisions and school districts in British Columbia is presented.

KEY WORDS: Age/sex population estimates; Small area; Residual net migration.

1. INTRODUCTION

The Central Statistics Bureau currently produces post-censal population estimates for a variety of sub-provincial areas using a regression approach (Central Statistics Bureau 1982). In addition to estimates of the total population by small area, age/sex specific estimates are also produced.

This paper outlines the method by which age/sex specific population estimates are derived for subprovincial areas of British Columbia, given an estimate of the total population.

2. OVERVIEW

The methodology used to derive the small area populations by sex and single years of age is divided into two parts.

The first part consists of examining historical residual net migration data compiled from censuses to derive a number of migration distributions by sex and single year of age for each small area (Shryock and Siegal 1980).

The second part of the methodology consists of aging the base population for each sex and adding births and subtracting deaths to yield a new population distribution for each area. This is referred to as the "natural base" population. The difference between the estimated total population by sex and the natural base population yields a residual term, which is equal to net migration by sex if the population and vital events for the two periods are exact. This small area sex specific residual term is distributed by single years of age according to a historical distribution, then added to the natural base population giving an age/sex specific population estimate for the area in the next time period.

Due to the timeliness of the input data, estimates of the total populations can be produced four months after the reference date of June 1, and the age/sex breakdowns one to two months later.

---

[1] Abridged version of the paper presented at the meeting of the Federal-Provincial Committee on Demography, Ottawa, November 28-29, 1985.

[2] D.S. O'Neil, SRL Sociometrics Resources Ltd., and C.D. McIntosh, Intersoft Resources Ltd., Central Statistics Bureau, Ministry of Industry and Small Business Department, Government of British Columbia, 2nd Floor, 1405 Douglas Street, Victoria, British Columbia, Canada V8W 3C1.
The views expressed in this paper are those of the authors and do not necessarily represent the views of the Government of British Columbia.

3. HISTORICAL NET MIGRATION DISTRIBUTIONS

Age/sex specific residual estimates of net migration were compiled for the census periods 1961/66, 1966/71 and 1971/76 for each of the 74 British Columbia school districts. These are referred to as the Historical Small Area Distributions.

Examination of these net migration distributions by small area showed them to be extremely unstable over time. In order to minimize the effects of this instability, a number of steps were taken.

First, migration distributions by small area were separated according to whether they occurred during a time of positive or negative total net migration. It was found that residual migration age distributions for many areas differed depending on whether net migration was positive or negative.

A further step taken to reduce the effects of unstable migration distributions was to group small areas of similar proportional migration distributions together, then calculate the positive and negative net migration distributions for each group of areas. These were called the Historical Grouped Distributions. Cluster analysis (using the SPSS/PC procedure) across selected age groups was used to group the historical small area migration distributions. Examination of cluster memberships from different periods resulted in the placing of the majority of areas into three clusters, while eight areas were maintained as unique independent clusters. Once areas had been arranged into groups, positive and negative migration distributions were calculated from the most recent periods of positive or negative net migration.

4. SMALL AREA POPULATION ESTIMATES BY SEX AND SINGLE YEAR OF AGE

As noted in Section 3, some areas showed considerable time-series variation in the residually calculated net migration distributions. This was likely the result of two factors. First, many of the areas under study possess small resource based economies subject to wide fluctuations, with consequent swings in migration levels. Second, a certain amount of instability is introduced when calculating a percentile distribution for a concept such as net migration, which may have either positive, negative, or zero values.

In order to guard against adopting a historical net migration distribution that may not be a representative distribution for the estimating year, five different historical sex-specific distributions were calculated, then distributed by single year of age. A description of these five different net migration distributions is given below.

1) The Historical Small Area Distribution for each small area having the same sign as the net migration to that small area was the first migration distribution.
2) The Historical Group Distribution for the group the small area belongs to, having the same sign as the net migration to that small area, was the second migration distribution.
3) The third migration distribution was calculated by separately totaling the migration from the most recent time period for all small areas with a positive and negative net migration, then calculating the age distributions.
4) The fourth distribution was the distribution of the natural base population for each small area.
5) The fifth and final distribution was the age distribution of migrants to British Columbia as a whole. For all the years under consideration, migration to B.C. has been positive, hence this is a positive distribution. Nevertheless, it was used as the fifth distribution regardless of whether the migration to a small area was positive or negative.

In some cases it was not possible to calculate all five distributions. This was the case if a small area never had a negative net migration in the past, but one is indicated for the estimating year under consideration. In situations such as this only distributions that can be calculated were used to distribute the small area net migration.

Empirical testing based on the 1981 Census indicated that of the five net migration distributions described above, number 1 (the Historical Small Area Distribution) produced the lowest average absolute percent error over all school districts and age groups, followed by number 2 (Historical Grouped Distribution), then number 3, etc. However, despite the fact that distribution number 1 produced the lowest error on average, it did not produce the lowest error in each case. Hence, a selection procedure was designed to substitute the population distribution produced by number 1, with either 2, 3, 4, or 5 in only those cases where the population distribution produced by number 1 was considered unrepresentative of the estimating year population distribution.

Empirical testing based on the 1981 Census resulted in the following selection procedure to be adopted.

First, all migration distributions possible were calculated and added to the natural base population, resulting in up to five possibilities for the small area estimated population by sex and single year of age in the next time period. These age/sex specific population estimates were then examined to determine which one produced the least change in the small area age structure from the previous year. This was done by first calculating the unweighted average percent difference between the age structures for each of the five possible populations in time t+1 to the population in time t. Next, the standard deviations about these averages were calculated, and the distribution with the lowest standard deviation is flagged. If the standard deviation produced by using the Historical Small Area Distribution was significantly greater than the smallest standard deviation (i.e. of the flagged distribution), then the Historical Small Area Distribution was rejected. This procedure was repeated with the Historical Grouped Distribution, and so on until one of the five possible populations was selected.

Once the "best" population in time $t+1$ was calculated for all small areas, two final adjustments were made. First, family allowance data was substituted for the age groups 0-14, and the populations for the rest of the age groups were pro-rated to keep the total population of each small area constant. The second adjustment was to pro-rate the population to ensure the age distribution of the sum of the small area population estimates was consistent with the British Columbia age distribution estimated by Statistics Canada.

5. EVALUATION OF THE CURRENT METHODOLOGY

The following tables summarize the error associated with the June 1, 1981 population estimates by five year age group to 70+, for 74 British Columbia school districts and 29 census divisions. The census division age/sex specific population estimates were derived by aggregating school district population estimates.

The accuracy of the small area age/sex specific population estimates derived from the previously described methodology was evaluated by producing 1981 population estimates by sex and 5 year age groups to 70+ for 74 school districts, then comparing these results to the 1981 Census. Two summary measures were used to evaluate the effectiveness of the age/sex specific population estimates. These were Average Absolute Percent Error (AAPE) and Index of Misallocation (IM). The AAPE is defined as:

$$AAPE = 100 \times \left[\sum_{i=1}^{N} \left| (P_{Ei} - P_{Ai})/P_{Ai} \right| \right]/N$$

where P_{Ei} is the estimated cell population for age group i, P_{Ai} is Census cell population for age group i, and N the number of cells. The IM is defined as:

$$IM = 100 \times \frac{1}{2}\left[\sum_{i=1}^{N}\left(\left|P_{Ai} - P_{Ei}\right|\right)\right] / \sum_{i=1}^{N} P_{Ai}$$

where P_{Ai} is the actual cell population for age group i, and P_{Ei} is the estimated cell population for age group i.

As seen in Table 1, relative to the 1981 Census the average absolute percent error over all age groups and regions is 6.20%, and the IM is 1.95%. The average percent errors for male and female are quite similar (AAPE's of 7.00% for both, and IM's of 2.15% for males and 2.08% for females).

By age, the highest errors occur in the 20-29 and 60-69 age groups. It should also be noted that there is some difference in the age distribution of errors between males and females. Males appear to have higher error in the upper age groups, while females have higher error in the very mobile 20-29 age groups.

Table 1

Error by Age Group Across School District
1981 Estimated Versus Census
Absolute Average Percent Error (AAPE) and Index of Misallocation (IM)

Age	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-4	3.33	0.96	3.94	1.21	3.62	1.04
5-9	2.80	0.76	3.28	0.88	3.62	1.02
10-14	2.33	0.64	3.54	0.84	2.88	0.87
15-19	5.20	2.01	5.68	2.01	6.18	2.24
20-24	13.32	4.77	13.50	4.62	14.54	5.12
25-29	8.31	4.07	8.42	3.70	9.41	4.65
30-34	5.02	2.12	5.42	2.45	5.72	2.06
35-39	4.88	1.33	5.73	1.62	5.38	1.34
40-44	4.52	1.33	5.84	1.51	4.67	1.52
45-49	3.60	1.22	4.47	1.37	4.78	1.49
50-54	5.66	1.33	5.86	1.48	6.68	1.54
55-59	6.11	1.72	6.19	1.78	7.82	1.97
60-64	8.86	2.44	10.35	2.95	8.91	2.17
65-69	10.60	2.66	12.53	3.52	11.44	2.30
70+	8.49	1.95	10.19	2.35	9.33	1.94
Average	6.20	1.95	7.00	2.15	7.00	2.08

As seen in Table 2 , on average higher percent errors are associated with areas of small population size. The higher percent errors in smaller areas may be associated with the instability of the smaller (resource based) economies, and associated instabilities in net migration distributions.

By census division, similar error patterns are observed. As seen in Table 3, the average absolute percent error across all regions and age groups is 4.83%, 5.19% for males and 5.60% for females. The IM is 1.27% for the total, 1.41% for males and 1.35% for females. Again, the error is bimodal, with peaks at 20- 29 and 60-69. In addition, the females have higher errors than males in the 20-29 age groups, while the reverse is true in the 60-69 age groups.

Table 2

School District Error by Population Size

Population Grouping	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-9,999	8.87	3.16	10.14	3.89	10.27	3.65
10,000-24,999	6.07	2.47	6.92	2.96	6.62	2.58
25,000+	3.66	1.67	3.92	1.78	4.09	1.78
School District Average	6.20	1.95	7.00	2.15	7.00	2.08

Table 3

Error by Age Group Across Census Division
1981 Estimated Versus Census

Age Group	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-4	2.37	0.54	3.20	0.76	2.28	0.58
5-9	1.52	0.50	1.71	0.55	2.13	0.68
10-14	1.69	0.39	2.75	0.57	2.50	0.60
15-19	3.81	1.39	3.79	1.30	4.68	1.63
20-24	9.83	3.07	9.30	2.91	10.90	3.41
25-29	7.02	3.04	7.30	2.87	8.09	3.37
30-34	3.28	1.29	3.31	1.43	3.85	1.25
35-39	3.34	0.66	3.06	0.57	4.21	0.88
40-44	3.86	0.88	4.29	1.01	4.16	0.90
45-49	2.91	0.70	3.20	0.75	3.75	0.83
50-54	4.82	0.64	4.41	0.75	6.10	0.86
55-59	5.49	1.34	5.36	1.55	6.94	1.30
60-64	7.88	1.95	8.37	2.29	7.94	1.74
65-69	8.48	1.89	10.30	2.67	9.79	1.43
70+	6.16	0.81	7.46	1.20	6.73	0.71
Avg	4.83	1.27	5.19	1.41	5.60	1.35

Table 4 (Census Division Error By Population Size) shows the improvement in error levels resulting from aggregating to larger sub-provincial areas. Table 7 illustrates the negative relationship between error levels and population size on a Census Division level.

A comparison of Tables 5 and 6 again demonstrates the improvement in error levels when aggregating to larger age/sex cell sizes. Although this does indicate that some precautions should be observed when utilizing age/sex estimates for some small areas, we do not believe it should preclude use of the estimates for these areas.

Table 4

Census Division Error by Population Size

Population Grouping	Total		Male		Female		N
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)	
0-39,000	7.22	1.94	7.55	2.13	8.79	2.29	10
40,000-59,999	4.32	1.82	5.03	2.14	4.91	1.83	10
60,000+	2.51	0.87	2.73	.98	2.84	0.90	9
Census Division Average	4.83	1.27	5.19	1.41	5.60	1.35	29

Table 5

- School District -
Number of Estimates by Error Range

	Average Absolute Percent Error Range				
	<5	5 to 10	10 to 15	15+	Total
No. of Cells	674	239	101	96	1110
Percent	61%	22%	9%	9%	100%

Table 6

- Census Division -
Number of Estimates by Error Range

	Average Absolute Percent Error Range				
	<5	5 to 10	10 to 15	15+	Total
No. of Cells	306	77	25	27	435
Percent	70%	18%	6%	6%	100%

6. FINAL REMARKS

The procedure outlined above has particular advantages for use in a region with well developed sources of historical small area population and vital statistics data. It is felt that a procedure utilizing net-migration estimates is relatively straightforward, produces acceptable error levels, and can produce age/sex estimates soon after the reference date. Although the optimal situation would be to have in- and out-migration estimates, currently little information is available on small area migration flows within British Columbia. One further improvement to the system being considered is the incorporation of Old Age Security counts to increase the stability and accuracy of estimates in the older age groups.

ACKNOWLEDGEMENTS

The authors would like to thank Don McRae, Steve Miller, Ravi Verma, Garnett Picot, and Paul Knapp whose input to and support for the development of the estimation breakdown system should not go unrecognized.

Table 7

Error by Census Division Across Age Groups
1981 Estimated Versus Census

Census Division	Total Population	Total		Male		Female	
		AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
1000 East Kootenay	53,725	4.24	2.04	5.24	2.29	3.88	2.15
3000 Central Kootenay	52,045	4.00	2.18	4.03	2.13	5.06	1.69
5000 Kootenay-Boundary	33,235	2.32	1.23	2.34	1.18	3.21	1.68
7000 Okanagan-Similkameen	57,185	5.04	2.64	6.02	3.08	4.72	2.49
9000 Fraser-Cheem	56,930	3.12	1.60	3.33	1.78	4.15	2.08
11000 Central Fraser Valley	115,015	3.14	1.43	3.46	1.52	3.65	1.81
13000 Dowdney-Alouette	62,000	2.10	1.15	2.56	1.23	2.26	1.32
15000 Greater Vancouver	1,168,700	1.63	0.94	1.68	0.93	1.67	0.98
17000 Capital	249,475	1.64	0.87	2.31	1.21	1.18	0.61
19000 Cowichen Valley	45,315	3.09	1.66	3.36	1.69	3.85	2.08
21000 Nanaimo	84,815	3.07	1.58	3.40	1.74	3.22	1.66
23000 Alberni-Clayoquot	32,560	2.75	1.36	2.88	1.27	3.27	1.68
25000 Comox-Strathcona	68,620	1.44	0.80	1.85	0.87	2.85	1.50
27000 Powell River	19,050	5.36	2.58	5.06	2.44	6.18	3.03
29000 Sunshine Coast	16,625	4.84	2.57	6.79	3.58	5.65	2.81
31000 Squamish-Liltooet	18,925	1.82	0.99	2.56	1.37	3.10	1.58
33000 Thompson-Nicola	102,430	2.13	1.10	2.07	0.10	2.65	1.37
35000 Central Okanagan	85,235	3.96	1.93	3.91	1.88	4.32	2.14
37000 North Okanagan	69,033	5.26	2.52	6.44	3.06	5.05	2.50
39000 Columbia-Shuswap	45,425	3.04	1.63	3.56	1.84	2.99	1.66
41000 Cariboo	58,810	3.18	1.93	3.90	2.18	3.42	2.06
43000 Mount Waddington	14,675	8.96	3.04	5.13	1.59	17.77	5.49
45000 Central Coast	3,050	17.99	7.62	21.62	8.86	14.92	7.34
47000 Skeena-Queen Charlotte	24,030	4.82	2.09	5.70	2.58	4.61	1.84
49000 Kitimat-Stikina	41,790	6.26	1.99	4.99	1.66	8.59	2.78
51000 Bulkley-Nechako	38,310	6.23	2.31	5.76	2.10	6.83	2.57
53000 Fraser-Fort George	89,430	3.50	1.41	3.39	1.25	3.72	1.68
55000 Peace River-Liard	55,340	8.00	2.95	9.43	3.65	7.34	2.83
57000 Stikine	2,685	17.15	6.89	17.89	6.88	22.39	8.35
Average Error		4.83	2.17	5.19	2.31	5.60	2.51

REFERENCES

CENTRAL STATISTICS BUREAU (1982). British Columbia municipal population estimation methodology. Unpublished report. Victoria: British Columbia Ministry of Industry and Small Business Development.

SPSS INC. (1984). *Statistical Package for the Social Sciences/PC.* Chicago, B265-B280.

SHRYOCK, H.S., and SIEGAL, J.S. (1980). The methods and materials of demography, 2. U.S. Bureau of the Census, 628-630.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 211-218
Statistics Canada

Estimating Population by Age and Sex for Census Divisions and Census Metropolitan Areas[1]

RAVI B.P. VERMA, K.G. BASAVARAJAPPA and ROSEMARY K. BENDER[2]

ABSTRACT

A methodology has been developed for producing population estimates by single years of age and sex for small areas (census divisions and census metropolitan areas). To assure reliability, the estimates by single years of age are grouped into five years and only these grouped data are recomended for dissemination. They are based on the age-sex composition of population from the last census, births by sex, deaths by single years of age and sex, estimates of migration by age and sex, and counts of family allowance recipients in the age group 1-14 years.

KEY WORDS: Cohort-component method; Mean absolute error; Index of dissimilarity; Separation factor.

1. INTRODUCTION

The objective of this paper is to describe the methodology for estimating population by age and sex for small areas (census divisions and census metropolitan areas), present findings of the evaluation of estimation methods, and finally to discuss the factors affecting the quality of estimates. According to the 1981 Census, the 266 Census divisions ranged in population from 2,000 to 2,000,000, and the 24 census metropolitan areas, from 100,000 to 3,000,000. The description of the estimation methods and principal data sources are presented in section 2. The results of the evaluation of migration and population estimates are given in section 3.

2. METHODOLOGY

The descriptions of the estimation methods, as well as the preparation of the basic input data are presented below.

2.1 Cohort-Component Method

For each census division (CD) and census metropolitan area (CMA), the cohort-component method is used to produce population estimates by age. The equations are as follows:

For the age 0, $$P_0^{t+1} = B - f_0 D_0 + \tfrac{1}{2} M_0 \quad (1)$$

For the age 1, $$P_1^{t+1} = P_0^t - \left[(1-f_0)D_0 + \tfrac{1}{2}D_1\right] + \tfrac{1}{2}(M_0 + M_1) \quad (2)$$

For ages 2 to 84, $$P_{a+1}^{t+1} = P_a^t - \tfrac{1}{2}(D_a + D_{a+1}) + \tfrac{1}{2}(M_a + M_{a+1}) \quad (3)$$

For ages 85+, $$P_{85+}^{t+1} = P_{84+}^t - \tfrac{1}{2}D_{84} - D_{85+} + \tfrac{1}{2}M_{84} + M_{85+} \quad (4)$$

---

[1] Revised version of the paper presented at the Federal-Provincial Committee on Demography meetings held on November 28-29, 1985 at Statistics Canada, Ottawa, Canada. This research was undertaken with support from the Small Area Data Program of Statistics Canada.

[2] Ravi B.P. Verma, K.G. Basavarajappa and Rosemary K. Bender, Demography Division, Statistics Canada,, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

where f_0 = Separation factor of deaths at age 0

M_a = Net migrants aged a between time t and $t+1$

B = Births between time t and $t+1$

D_a = Deaths at age a between time t and $t+1$

P_a^t = Population aged a at time t.

The cohort-component method is also used at the provincial level by Statistics Canada (Statistics Canada, Catalogue No. 91-210), and by the province of British Columbia for producing population estimates by age at the census division, school and health district levels (Central Statistics Bureau 1980).

2.2 Preparation of Basic Input Data

Since we are proposing to produce preliminary postcensal population estimates within eight months after the reference date, final data on components of population change cannot be used because they do not become available until after 18 to 24 months. Consequently, estimates would have to be used for each component.

Births and Deaths

Preliminary estimates of births by sex for year (t) are obtained by multiplying the proportional distribution by small areas of provincial total births by sex for year ($t-1$) with the provincial preliminary total births for year (t). Similarly, preliminary estimates of deaths by age and sex for year (t) are obtained by multiplying the proportional distribution by small areas of provincial total deaths by age and sex for year ($t-1$) with the provincial preliminary total deaths for year (t). Finally, they are converted into cohort deaths on the assumption that dates of birth of those who die and the number of deaths are uniformly distributed over a 12 month period except for deaths of age 0. The formulae are as follows:

For age 0,
  Cohort deaths (0) = deaths (0) × 0.89
For age 1,
  Cohort deaths (1) = [deaths (0) × 0.11] + [deaths (1) × 0.5]
For ages 2 to 84,
  Cohort deaths (age) = [deaths (age-1) × 0.5] + [deaths (age) × 0.5]
  Cohort deaths (85+) = deaths (84) × 0.5 + deaths (85+).

In the above formulae, the separation factors (f) are 0.89 for age 0, 0.11 for age 1 and 0.5 for all other ages.

Residual Net Migration

First, the estimates of total population for the postcensal years for CDs and CMAs prepared by the regression-nested procedure are split by sex using the sex composition from the latest census. The regression-nested procedure is described elsewhere (Statistics Canada, Catalogue No. 91-211). For males and females, residual total net migration is computed by taking the difference between the population change and the natural increase. For each area, this is distributed by five year age groups using migration data by age from three sources: residual

net migration from the 1976 and 1981 censuses, migration data from income tax files and the 1981 mobility question. The mobility question referred to is "Where were you on June 1, 1976?" in the 1981 Census. From the responses obtained for this question, in-migrants to and out-migrants from each small area can be tabulated. The five year age groups are split into single years of age using SPRAGUE multipliers. Before applying Sprague multipliers, the residual net migration is first split into in and out migration. Using in and out tax migration data as a reference, this calculation is done individually for each five-year age group.

$$\text{Residual In-Migration} = \frac{\text{Tax Data In-Migration}}{\text{Tax Data Net Migration}} \times \text{Residual Net Migration}$$

$$\text{Residual Out-Migration} = \text{Residual In-Migration} - \text{Residual Net Migration}$$

Using the preceding ratios, major problems occur when the split net migration is not of the same sign as the reference tax data on net migration. In this case, the sign of the split net migration is kept, but the resulting in and out migration are exchanged to yield the appropriate sign. This is based on the assumption of equal magnitude of a reversal of the migration flow.

2.3 Counts From The Family Allowances File, Ages 1-14 years

Estimates of population produced by the cohort-component method for the age groups 1-4, 5-9, 10-14 are replaced by counts of family allowance recipients at these ages which are readily available for CDs and CMAs, within 3 to 4 months after the reference date. Family allowances are paid universally in Canada and hence the counts are considered to be complete for all practical purposes. The data on the family allowance recipients are not provided by sex. Hence they are split into males and females using the sex composition from the latest census.

2.4 Adjustments for Consistency with Provincial and Census Division Estimates

Postcensal regression-nested estimate of total population of each CD and CMA become available within six months after the reference date. In addition, provincial estimates of population also become available by age and sex about the same time. Estimates of population by age and sex prepared as described above for the CDs within each province are controlled with respect to the census division total population estimates, and to the provincial population estimates by age and sex on a pro rata basis. For the census metropolitan areas, the age and sex totals are adjusted only to the CMA total population estimate.

3. EVALUATION

The evaluation is done with respect to three criteria: (i) accuracy; (ii) timeliness and (iii) consistency. Each of these is discussed below.

3.1 Accuracy

The accuracy of population estimates by age and sex depends to a large extent on the accuracy of estimation of the age-sex distribution of migrants, as the data on deaths by age and sex are considered satisfactory. Thus an evaluation of population estimates by age and sex indirectly throws light on the accuracy of migration estimates by age and sex. The accuracy is examined by comparing the estimates with the corresponding census counts.

Table 1

Distribution of Census Divisions/CMAs Showing the Accuracy of Population Estimates by Age, 1981

Provinces	Methods of Migration Estimation	Levels of Mean Absolute Error (%) by Sex							
		Males				Females			
		Under 3	3-5	5-10	10+	Under 3	3-5	5-10	10+
Newfoundland	R	8	2	0	0	10	0	0	0
	M	2	7	1	0	3	3	3	1
	T	0	0	5	5	1	1	3	5
Prince Edward Island	R	1	1	1	0	2	0	0	1
	M	1	2	0	0	1	2	0	0
	T	0	0	1	2	0	0	2	1
Nova Scotia	R	8	4	3	3	8	5	3	2
	M	3	6	4	5	5	6	2	5
	T	0	2	8	8	1	3	11	3
New Brunswick	R	10	1	2	2	7	4	3	1
	M	4	7	3	1	3	8	3	1
	T	0	2	5	8	0	4	5	6
Quebec	R	13	27	23	13	24	18	19	15
	M	12	26	26	12	17	23	21	15
	T	1	5	37	33	3	13	32	28
Ontario	R	30	8	8	7	37	5	4	7
	M	8	16	18	11	21	10	10	12
	T	0	8	34	11	4	10	31	8
Manitoba	R	4	6	8	5	4	6	8	5
	M	1	5	12	5	0	6	7	10
	T	0	1	8	14	0	1	4	18
Saskatchewan	R	10	5	1	2	9	5	2	2
	M	1	11	5	1	4	5	5	4
	T	1	1	10	6	1	1	12	4
Alberta	R	9	3	1	2	8	3	3	1
	M	5	5	3	2	5	4	5	1
	T	0	2	5	8	0	3	5	7
British Columbia	R	19	3	3	4	23	1	1	4
	M	9	13	2	5	14	8	3	4
	T	0	0	13	16	0	3	16	10
CMA	R	14	8	2	0	19	3	2	0
	M	2	17	5	0	10	9	4	1
	T	1	7	13	3	1	10	12	1

Note: R: Residual based age distribution of migrants, 1976-81.
M: Mobility based age distribution of migrants, 1981.
T: Annual tax migration data.
Source: Demography Division, Statistics Canada, 1985.

For each CD and CMA, three sets of population estimates by age and sex as of June 1, 1981 produced by using the age distribution of migrants from the three sources (residual (1976-81), mobility (1976-1981) and annual tax files) and counts from family allowance files as described in sections 2.1 to 2.4 were compared with the 1981 census counts. The differences were termed errors and for each small area, a summary index known as the "mean absolute error" (MAE) was computed by taking the arithmetic mean of percentage errors disregarding

the sign for 16 five year age groups. The smaller the value of this index, the more accurate are the estimates. In Table 1, a classification of CDs by provinces and of CMAs is presented for four levels of mean absolute error: under 3%, 3-5%, 5-10% and over 10%. Overall, it appears that the residual based age distribution of migrants gives better estimates. For males, about 66% of the total number of census divisions had an MAE under 5%. For females this percentage was slightly higher, at 69%. In contrast, lower percentages were observed for the mobility (55% and 57%) and tax migration data (9% and 19%), for males and females, respectively.

For CMAs too, the residual age distribution of migrants seems to give better estimates. The proportions of cases with MAE under 3% were 58% and 79% for males and females respectively. Mobility and tax based age distributions of migrants ranked second and third respectively, for both males and females.

With the exception of Prince Edward Island, the relative accuracy of the three sets of age distribution observed for Canada largely holds good for each province. This is true for both males and females. However, in some cases the residual based age distributions seem to give results similar to the mobility based distributions. Such similarity was observed for males in three provinces (Newfoundland, New Brunswick and British Columbia), whereas for females it was found only in New Brunswick.

It should be noted that the age distribution of migrants derived by the residual method uses the census age distributions of 1976 and 1981. Consequently, the population estimates as of June 3, 1981 prepared by using the migrant age distribution based on the residual method can be expected to be similar to the 1981 census age distribution. Hence, on the basis of this comparison we cannot conclude that the migrant age distribution derived by the residual method is better than the distribution derived from mobility question or from tax files.

Table 2 presents the percentage distribution of CD and CMA outliers. The outliers are those CDs with an MAE of over 10% and those CMAs over 5%. They are presented by sex and the three sources of migrant age distributions. As expected, both for males and females, the proportion of outliers is generally low for estimates using residual based age distribution. On the other hand, the percentage of outliers tends to be high for estimates using tax based migration distribution.

Temporal Stability of the Three Sets of Estimates During Postcensal Years, 1982-1984

For postcensal years, as there are no standard age distributions with which the estimates can be compared, the three population estimates by age and sex are compared with each other to learn of the temporal stability among them. A summary index knows as the "index of dissimilarity" calculated as half of the sum of absolute differences in two percentage age distributions is used for this purpose. The range of the index is from 0 to 100. The smaller the value, the greater is the similarity between the two distributions compared. The small areas are classified into three levels of dissimilarity: (i) the smallest level of difference with indices between 0% and 5%; (ii) the medium level of difference with indices between 5 to 10% and (iii) the outliers showing the index value of 10% and over. The classification of CDs is presented in Table 3 and that of CMAs in Table 4.

From Table 3, it appears that all the three population distributions tend to be similar and on average, a high percentage of cases, about 90%, are in the smallest category of differences (0%-5%) with only about 7% falling in the 5% to 10% category.

The percentage of cases with the extreme level of differences (index of dissimilarity exceeding 10%) were also examined for the ten provinces and their total. For males, the percentages of extreme cases were small, 3 to 5% between the residual and mobility based age distributions. For females, a relatively higher proportions of outliers were noticed. For other comparisons, residual vs tax based, and mobility vs tax based, slightly higher proportions of outliers were found. The results were similar for census metropolitan areas (see Table 4).

Table 2

Percentage of Outliers[a] Among Census Divisions by Province, and of CMA's 1981

Provinces	Males			Females		
	R	M	T	R	M	T
Newfoundland	0	0	50	0	10	50
Prince Edward Island	0	0	67	33	0	33
Nova Scotia	17	28	44	11	28	17
New Brunswick	13	7	53	7	7	40
Quebec	17	16	43	20	20	37
Ontario	13	21	21	13	23	15
Manitoba	22	22	61	22	43	78
Saskatchewan	11	6	33	11	22	22
Alberta	13	13	53	7	7	47
British Columbia	14	17	55	14	14	34
Total	15	16	43	15	20	35
CMA	8	21	67	8	21	54

Note: R: Residual based age distribution of migrants.
M: Mobility based age distribution of migrants.
T: Tax based age distribution of migrants.
[a] The outliers are those CDs with MAE of over 10% and those CMAs with MAE of over 5%.
Source: Table 1.

Table 3

Distribution of Census Divisions by Level of Index of Dissimilarity Obtained by Comparing the Age Distributions of Population Based on Residual, Mobility and Tax Migration Sources, 1982 to 1984

Year/ Index of Dissimilarity	Males			Females		
	Residual vs Mobility	Residual vs Tax	Mobility vs Tax	Residual vs Mobility	Residual vs Tax	Mobility vs Tax
YEAR 1982						
0-5	245	237	242	240	241	234
5-10	7	13	10	8	5	11
10+	8	10	8	12	14	15
Total	260	260	260	260	260	260
YEAR 1983						
0-5	235	221	223	230	229	223
5-10	11	18	21	10	13	13
10+	14	21	16	20	18	24
Total	260	260	260	260	260	260
YEAR 1984						
0-5	240	226	229	235	233	231
5-10	11	16	14	15	13	12
10+	9	18	17	10	14	17
Total	260	260	260	260	260	260

Source: Demography Division, Statistics Canada, October 1985.

Table 4

Distribution of Census Metropolitan Areas by Level of Index of Dissimilarity Obtained by Comparing the Age Distributions of Population Based on Residual, Mobility and Tax Migration Sources, 1982 to 1984

Year/	Males			Females		
Index of Dissimilarity	Residual vs Mobility	Residual vs Tax	Mobility vs Tax	Residual vs Mobility	Residual vs Tax	Mobility vs Tax
YEAR 1982						
0-3	24	24	24	23	22	23
3-5	0	0	0	0	1	0
5+	0	0	0	1	1	1
Total	24	24	24	24	24	24
YEAR 1983						
0-3	22	23	22	21	20	20
3-5	2	0	0	1	2	3
5+	0	1	2	2	2	1
Total	24	24	24	24	24	24
YEAR 1984						
0-3	21	21	21	21	20	20
3-5	2	2	2	0	1	0
5+	1	1	1	3	3	4
Total	24	24	24	24	24	24

Source: Demography Division, Statistics Canada, October 1985.

In conclusion, it may be said that although the three age distributions of migrants (residual, mobility and tax based) differed from each other, age distributions of population resulting from these were largely similar.

3.2 Timeliness

Timeliness refers to the availability of estimates within as short a time as possible after the reference date. Using the preliminary population totals (regression-nested estimates) which become available within six months from the reference date, the estimated numbers of births, deaths by age and net migrants by age as described in Sections 2.1 to 2.4, the population estimates by age and sex for CDs and CMAs could be prepared within eight months of the reference date.

3.3 Consistency

Consistency refers to the consistency in the sources of data sets used for estimation at various levels of administrative or other disaggregated areas and to the uniformity in the methods of estimation. While in certain cases, a different method may have to be used, it is highly desirable to use the same method throughout in order to ensure the methodological consistency of various levels of geographic disaggregation.

For provinces, CDs and CMAs, the sources of data are the same for births and deaths: the vital registration records. For migration data too, the sources are the same namely, tax files and mobility data from the census for all levels of geographic disaggregation. However, an additional data set, the residual age data derived from the two consecutive censuses is also used.

There is full methodological consistency between provinces and other levels as the cohort-component method is used in all cases.

4. CONCLUSION

By using the cohort-component method, three sets of estimates by age and sex have been prepared for CDs and CMAs. Each set uses a different migration component by age and sex: (i) tax file based; (ii) mobility data from the latest census and (iii) the residual derived from the two consecutive censuses.

Although the three age distributions of migrants differ from each other, the resulting estimates of population by age and sex were largely similar. Each set involves its own assumptions. Using a residual age distribution of migrants for postcensal estimation assumes that the age distribution remains constant for the period of estimation. A similar assumption is involved in using mobility data by age and sex for postcensal years. The data from tax files assume that the age-sex distribution remains the same for any two consecutive years. However, the type of movement measured by each of these sets is not the same. The residual measures only the net movement between the two consecutive censuses (e.g. 1976-81). The mobility question also measures five-year movements ranging from 0-4 years. The tax files, on the other hand show the movement during roughly a 12 month period. On the basis of the comparisons made in the paper, it cannot be concluded that one migrant data set giving rise to population estimates is better than the other. A more satisfactory evaluation of the three sets of estimates can be made only when the next census results become available.

REFERENCES

CENTRAL STATISTICS BUREAU (1980). Population estimates by age groups for school districts, 1977-80. Unpublished Technical Report, Government of B.C.

NORRIS, D., and STANDISH, L. (1983). A Technical report on the development of migration data from taxation records. Technical Report, Administrative Data Development Division, Statistics Canada.

STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada and the Provinces*, Vol. 2, 2nd issue. Catalogue 91-210, Ottawa: Ministry of Supply and Services.

STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Regression-Method)*. Catalogue 91-211, Ottawa: Ministry of Supply and Services Canada.

STONE, Leroy O. (1980). Evaluating the relative accuracy and significance of net migration estimates. *Demography*, 4, 310-330.

Survey Methodology, December 1985
Vol. 11, No. 2, pp. 219-222
Statistics Canada

Experience with Small Area Population Estimates[1]

ROSEMARY K. BENDER[2]

ABSTRACT

Statistics Canada's current methodologies forestimating the population of census divisions and census metropolitan areas are the regression-nested and component methods. This paper presents the experience with these estimates for the period 1981 to 1985, focusing on problems encountered with the input data on family allowance recipients.

KEY WORDS: Regression-nested estimates; Component estimates; Family allowance recipients; Postal code files.

1. INTRODUCTION

Statistics Canada's current methodologies for estimating the population of census divisions (CDs) and census metropolitan areas (CMAs) are the regression-nested and component methods. The regression estimates for 1982, 1983 and 1985 were published in Catalogue No. 91-211 on schedule. Those for 1984 were only made available in March of 1985. There was a delay in obtaining the input data on family allowance. Furthermore, as explained below, we encountered problems with the quality of these data. In particular, the resulting population estimates for CMAs were not acceptable and an alternate methodology had to be used.

Component estimates of the population for CDs and CMAs have been published in Catalogue No. 91-212 on schedule for 1982 and 1983. We should release the 1984 estimates by April 1986. An evaluation of the component estimates produced thus far has shown the data to be of good quality.

2. ADJUSTMENTS

Since introducing the regression estimates for CDs and CMAs in 1982, some adjustments to the data and the methodology have been necessary. They are summarized below:

- For the 1983 estimates for the CD Chicoutimi and the CMA Chicoutimi- Jonquière in the province of Quebec, the family allowance data was adjusted based on the growth pattern of the previous year. The problem was traced to postal codes used to obtain the family allowance data.
- In 1984, 17 census divisions estimates were imputed with preliminary component estimates.
- In 1984, we decided to publish for the CMA of Calgary, estimates based on the annual census conducted by the city. This will be done for the entire 1981-1986 period.
- In 1984, we developed a new methodology for all CMAs other than Calgary, which aggregates census division regression estimates. This will be used for the entire 1981-1986 period.

The following sections explain the problems encountered in more detail.

---

[1] Abridged version of the paper presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa, Canada.

[2] Rosemary K. Bender, Demography Division, Census and Demographic Statistics Branch, Statistics Canada, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

3. PROBLEMS WITH INPUT DATA FOR REGRESSION ESTIMATES

There was a delay in producing 1984 estimates due to problems encountered in obtaining data on family allowance recipients from Health and Welfare Canada, and the appropriate postal code translation files necessary to process these data.

i) *Family Allowance Data*

The numbers of Family Allowance recipients as of June 1, is generally available by mid September of each year. The 1984 data from Health and Welfare Canada however, were delayed as a result of decentralization of the regional operations of the program in Ontario. Problems were also encoutered in the files of all provinces with respect to information on effective dates of transfer and reason codes for inter-area transfers. The 1984 data were released to Statistics Canada in an unedited form in November. Corrective actions were taken by Health and Welfare Canada, and Family Allowance data as of June 1, 1985 was on schedule.

ii) *Postal Code Files*

The data on family allowance recipients from Health and Welfare Canada is coded by postal code. Therefore, to identify the children receiving family allowance in each CD and CMA, a file must be created that groups the postal codes by CD and CMA. This is done using a master file that contains all the postal codes in Canada, with detailed geographic codes that are used to assign the postal codes to any level of geographic disaggregation.

Problems have arisen that were unexpected and in some cases had serious consequences. For our estimates, it is important that the postal code files used each year by Health and Welfare Canada be consistent with the one that was used to develop the regression model. The only change in the file should be the addition of new postal codes. Any shifting of postal codes from one region to another can result in changes to the population that do not actually occur.

The problems we encountered stem from the fact that since we developed our regression model, different divisions and departments have produced the postal code files. In 1982 and 1983, it was done by the Administrative Data Development Division of Statistics Canada. In 1984, the Standards Division of Statistics Canada took over the responsability and in 1985 it was done by Health and Welfare Canada. Each had its own approach resulting in family allowance data that was not consistent from year to year. Two different types of problems arose. We have resolved the first. However, the second will persist throughout the 1981-1986 postcensal period.

The first source of difficulty was the shifting of postal codes from one area to another. The master file is created by the Standards Division of Statistics Canada. However, in some cases, the CD or CMA geographic code is blank or wrong. For CDs this occurs mostly with rural codes, where postal codes often refer to post offices covering large territories across CD boundaries. The inclusion of the CMA geographic codes is fairly recent, and the quality improves each year. Thus, our initial assumption that the postal code file would be consistent from year to year was not quite true. There are changes made each year.

Our files were initially created by the Administrative Data Development Division (ADDD) of Statistics Canada. They made changes in their copy of the master file before proceeding to group the data. In 1984 the Standards Division took over producing our file. When we became aware of the consequences this would have, we developed with ADDD a way to match the original master file with the latest master file from Standards Division, adding only the new postal codes. Any changes to the CD or CMA codes were ignored. We realise that by doing this we do not have the most accurate postal code file available. However, for our

purposes, we are interested in the changes to the proportions of children receiving family allowance. The effect of using some erroneous, but consistent postal codes is that we include or exclude some children from another area in the calculation of proportions. The proportions would not be significantly different from those using correct postal codes, but would change if these children were suddenly excluded or included.

This process of adding only new codes to our postal code file improved significantly the quality of the 1984 family allowance data for census divisions. Only 17 of the 231 regression estimates of CDs (excluding those of British Columbia, as they produce their own regression estimates) needed to be imputed. Because of the delay in obtaining the data, we were able to use preliminary estimates from the component method. For census metropolitan areas, there were still inconsistencies, which we beleive are due to a different type of problem.

When the postal codes are grouped by CDs and CMAs, they are also converted into ranges of postal codes. For example, if the postal codes A1A1A1, A1A1A2, A1A1A3 and A1A1A4 all have the same CMA code, then they will be combined into the range A1A1A1-A1A1A4. However, in processing the over 600,000 postal codes, certain assumptions are made, depending on the software. If, in the above example A1A1A2 was not there, the program may still create the same range, assuming that if A1A1A2 did exist, it would have the same CMA code as the others in the range. This type of assumption could alter the family allowance data processed for each region. Furthermore, if different softwares are used each year, serious inconsistencies can arise.

We believe this is the major cause for the poor quality in the family allowance data for CMAs. The softwares used by the ADDD and Standards Divisions were different. What complicated matters even more was that as of 1985, the entire operation is now done by Health and Welfare Canada, again using a different software. We therefore had to disregard the data and develop an alternative methodology for CMAs.

4. METHODOLOGICAL CHANGE FOR CMAs

The CMA estimates previously released for 1982 and 1983 were based on the same regression-nested procedures as for census divisions. In the evaluation of the 1984 estimates, however, estimates for many census metropolitan areas were found to be inconsistent with alternate sources and past growth trends. As described above, the problems seem more related to the quality of the input files rather than to methodology.

Taking into account these inconsistencies as well as comments from the provincial focal points, it was decided to use an alternate methodology. This new methodology was previously developed for estimating various CMA components of population change. It consists of aggregating census divisions regression estimates, using the ratio of the population of the CMA to that of overlapping CDs, as observed the previous year by the component method. In comparing estimates for 1981, obtained through this methodology, with the 1981 Census counts for census metropolitan areas, an average absolute error of 1.3% as observed, as compared to 2.3% for the previous methodology.

To maintain consistency in methodology for the entire 1981-1986 period, the alternate method has been used to derive the CMA estimates for 1982 to 1985, and will be used for 1986. That is, estimates of population for CMA's other than Calgary are obtained by aggregating the census division regression-nested estimates, and those for Calgary as described below, are based on the annual census conducted by the city.

In 1984, it was found that the regression-nested estimates for Calgary CMA for 1982 and 1983 were too high in comparison with the census counts conducted annually by the city of Calgary. The component estimates also supported the idea of adjusting the regression-nested estimates for Calgary. It was decided to publish estimates based on the city of Calgary

census count extrapolating the April data to June 1. This is in line with Statistics Canada policy where, when there is a complete enumeration, this should be considered over an estimate prepared by an indirect procedure, unless there is evidence that the enumerated count is suspect.

5. COMPARISON WITH OTHER DATA SOURCES

The regression and component estimates are compared with alternative data sources whenever possible. We receive from the Saskatchewan and Alberta governments the number of people registered in their respective health care programs. These data are used in the regression model. However, they are also evaluated for consistency with the family allowance data and past growth trends. In most cases they were consistent, and differences were traced to the problems encountered with family allowance data.

The Quebec Bureau of Statistics produces annual population estimates of their administrative regions which are subdivisions of the Quebec CDs. Their data are comparable to ours except for the CD of Nouveau Québec. This census division, located in northern Quebec, is largely comprised of unorganized territories, and it is difficult to estimate the population. The BSQ generally adopts our estimates, though for 1984 it imputed its own estimate for Nouveau Québec.

We also appreciate feedback from users who may have access to specific local area data.

6. CONCLUSION

The methods used to produce population estimates for census divisions and census metropolitan areas have in general functionned very well. However, in the case of the regression estimates, problems with input data made it necessary to impute estimates for certain CDs with alternate data, and to revise the methodology for CMAs.

The problems encountered were mostly related to the family allowance data and the postal code files that are necessary to process these data. Most of the problems have been resolved. However, as Health and Welfare are now taking over the responsibility of creating the postal code files, the 1986 data may still have problems of consistency and will have to be carefully evaluated.

Despite these problems, the regression methodology with certain adaptations will be used to produce estimates for 1986. If, however, we decide to continue with the methodology for the 1986-1991 period, we must first ensure that consistent postal code files be processed by the same department throughout the period.

EDITORIAL COLLABORATORS

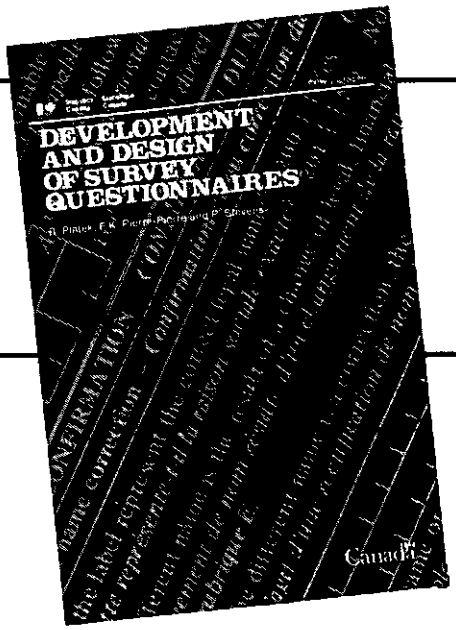
Following the Journal's policy, all papers published in Volume 11 (issues 1 and 2) were refereed except the paper by M. Wilk and the papers selected from those presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa. (Condensed version of these selected papers are included in this issue to provide the readers with additional information on recent methodological developments in an important area of applications.)

The Survey Methodology Journal wishes to thank the following persons who have served as referees during the past year.

D.A. Binder
Y.P. Chaubey
G.H. Choudhry
E.B. Dagum
J. Gambino
G.B. Gray
M.A. Hidiroglou
S.K. McKenzie
M. Morry
D.A. Pierce
B. Quenneville
C.E. Särndal
A. Satin
A. Singh
J. Tourigny
A. Van Barren

Statistics Canada Statistique Canada

DEVELOPMENT AND DESIGN OF SURVEY QUESTIONNAIRES



Successful implementation of surveys depends to a large extent on good questionnaire design. But the failure to devote sufficient attention, care and resources to questionnaire development is surprisingly common in current survey practice. As a consequence, many surveys fail to achieve their full potential.

From its Introductory comment through its four chapters dealing with organizing to design, development, production and evaluation of questionnaires, *Development and Design of Survey Questionnaires* intends to promote good questionnaire design and serve as a reference and training tool. It provides a discussion of issues related to wording, format, layout, etc. illustrating these with examples from recent federal surveys. A Checklist and Bibliography complete the 119 page text. (15 cm. x 23 cm.)

Contents

ORDER FORM

PF 02922

Mail to:
Publications Sales and Services
Statistics Canada
Ottawa, K1A 0T6

(Please print)

Company:
Dept.:
Attention:
Address:
City: Tel.:
Province: Postal Code:

☐ Purchase Order Number (Please enclose)

☐ Payment enclosed $

CHARGE TO MY:

☐ MASTERCARD ☐ VISA ☐ Statistics Canada

Account No.:

Expiry date

☐ Bill me later

My client reference number is:

Signature:

Catalogue No.	Title	Quantity	Price	Total
12-519E	**Development and Design of Survey Questionnaires**		**$25 in Canada $26 other countries**	

Cheques or money orders should be made payable to the Receiver General for Canada/Publications, in Canadian funds or equivalent.

Canada

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.