C-4

SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1993
•
VOLUME 19
•
NUMBER 2

Statistics Canada Statistique Canada

Canadä

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1993 • VOLUME 19 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

December 1993

Price: Canada: $45.00
United States: US$50.00
Other Countries: US$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal of Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 19, Number 2, December 1993

CONTENTS

In This Issue

Papers covering a variety of topics are included in this issue of *Survey Methodology*. In the first paper, Biemer and Atkinson present a general methodology for constructing and evaluating model prediction estimators of measurement bias for a stratified two-phase design with simple random sampling in each phase. For evaluation, they extended the bootstrap methodology of Bickel and Freedman to two-phase sampling. The example used for illustration indicates that improvements over the traditional net difference estimator and thus savings in the cost of reinterview surveys are possible.

The paper by Armstrong and Mayda was originally intended for the special section *Record Linkage and Statistical Matching*. The authors consider model based estimation of classification error rates in record linkage. The class of models considered allows for non-independence of match status of different matching fields within a record pair. Estimation methods are developed and different methods of error rate estimation are compared using both synthetic and real data.

Pfeffermann and Bleuer consider estimation for small areas using data from a rotating panel survey over time. Their approach is model based, with a state space model for the population values over time and separate autoregressive models for the survey error series from each panel. To achieve a measure of robustness, the small area estimators are further constrained to add up to direct survey estimators within pre-defined larger areas. The approach is demonstrated using Canadian Labour Force Survey data for the Atlantic provinces.

Mian and Laniel discuss two iterative procedures to find the maximum likelihood estimates of a non-linear benchmarking model that seems suitable for economic time series from large sample surveys. Closed form expressions for the asymptotic variances and covariances of the benchmarked series and of the fitted values are also provided. The methodology is illustrated using Canadian retail trade data.

Deville uses superpopulation models to anticipate, before data collection, the variances of estimates of ratios. Based on models that are both simple and realistic, he produces expressions of varying complexity and then optimizes them. He deals with the problem of estimating the frequency of errors in the population of forms collected during the quality control of the French census.

Asymptotic techniques are used by Casady and Valliant to study post-stratification from a design-based, conditional point of view. The authors derive the large sample bias and mean squared error of the standard post-stratified estimator, the Horvitz-Thompson estimator, a ratio estimator and a new post-stratified regression estimator. The developed theory is empirically tested using real and artificial populations. The problem of bias due to defective frames is also addressed.

Bandyopadhyay and Adhikari study estimation based on frames where some units are listed more than once, each time with a different identification. The mean square errors of estimators from imperfect and perfect frames are compared. Estimation of a population ratio, mean and total when no auxiliary information is available on the frame is considered.

Roesch, Green and Scott present a generalized concept for all of the commonly used methods of forest sampling. The concept views the forest as a two-dimensional picture which is cut up into pieces like a jigsaw puzzle, with the pieces defined by the individual selection probabilities of the trees in the forest.

The paper by Kalton and Citro is a revised version of the keynote address given at the Statistics Canada Symposium 92 on longitudinal surveys. The paper discusses how different designs for surveys over time satisfy various analytic objectives. The author then concentrates on panel surveys and talks about decisions that need to made when designing them.

The Editor

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 127–136
Statistics Canada

Estimation of Measurement Bias Using a Model Prediction Approach

PAUL P. BIEMER and DALE ATKINSON[1]

ABSTRACT

Methods for estimating response bias in surveys require "unbiased" remeasurements for at least a subsample of observations. The usual estimator of response bias is the difference between the mean of the original observations and the mean of the unbiased observations. In this article, we explore a number of alternative estimators of response bias derived from a model prediction approach. The assumed sampling design is a stratified two-phase design implementing simple random sampling in each phase. We assume that the characteristic, y, is observed for each unit selected in phase 1 while the true value of the characteristic, μ, is obtained for each unit in the subsample selected at phase 2. We further assume that an auxiliary variable x is known for each unit in the phase 1 sample and that the population total of x is known. A number of models relating y, μ and x are assumed which yield alternative estimators of $E(y - \mu)$, the response bias. The estimators are evaluated using a bootstrap procedure for estimating variance, bias, and mean squared error. Our bootstrap procedure is an extension of the Bickel-Freedman single phase method to the case of a stratified two-phase design. As an illustration, the methodology is applied to data from the National Agricultural Statistics Service reinterview program. For these data, we show that the usual difference estimator is outperformed by the model-assisted estimator suggested by Särndal, Swensson and Wretman (1991), thus indicating that improvements over the traditional estimator are possible using the model prediction approach.

KEY WORDS: Reinterview; Repeated measures; Response error; Bootstrap.

1. INTRODUCTION

It is well-known in the survey literature that when responses are obtained from respondents in sample surveys, the observed values of measured characteristics may differ markedly from the true values of the characteristics. Evidence of these so-called measurement errors in surveys has been collected in a number of ways. For example, the recorded response may be checked for accuracy against administrative records or legal documents within which the true (or at least a more accurate) value of the characteristic is contained. An alternative approach relies on revised reports from respondents via reinterviews. In a reinterview, a respondent is recontacted for the purpose of conducting a second interview regarding the same characteristics measured in the first interview. Rather than simply repeating the original questions in the interview, there may be extensive probes designed to elicit more accurate responses, or the respondent may be instructed to consult written records for the "book values" of the characteristics. For some reinterview surveys, descrepancies between the first and second interviews are reconciled with the respondent until the interviewer is satisfied that a correct answer has been obtained. Forsman and Schreiner (1991) provide an overview of the literature for these types of reinterviews. Other means of checking the accuracy of survey responses include: (a) comparing the survey statistics (*i.e.*, means, totals, proportions, *etc.*) to statistics from external sources that are more accurate; (b) using experimental designs to estimate the effects on survey estimates of interviewers and other survey personnel; and (c) checking the results within the same survey for internal consistency.

The focus of the current work is on estimators of measurement bias from data collected in true value remeasurement studies, *i.e.*, record check and reinterview studies, where the objective is to obtain the true value of the characteristic at, perhaps, a much greater cost per measurement than that of the original observation.

Because of the high costs typically involved in conducting reinterview studies, repeated measurements are usually obtained for only a small fraction of the original survey sample. While the sample size may be quite adequate for estimating biases at the national and regional levels, they may not be adequate for estimating the error associated with small subpopulations or rare survey characteristics. In this paper, our objective is to consider estimators of response bias having better mean squared error properties than the traditional estimators. The basic idea behind our approach can be described as follows.

In a typical remeasurement study, a random subsample of the survey respondents is selected and, through some means, the true values of the characteristics of interest are ascertained. Let n_1 denote the number of respondents to

[1] Paul P. Biemer, Principal Scientist, Center for Survey Research, Research Triangle Institute, Research Triangle Park, NC 27709; Dale Atkinson, Supervisory Mathematical Statistician, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Va 22030.

the first survey and let n_2 denote the number selected for the subsample or evaluation sample. The usual estimator of response bias is the net difference rate, computed for the n_2 respondents in the evaluation sample as

$$\text{NDR} = \bar{y}_2 - \bar{\mu}_2, \tag{1.1}$$

where $\bar{y}_2$ is the sample mean of original responses and $\bar{\mu}_2$ is the sample mean of the true measurements. A disadvantage of the NDR is that it excludes information on the $n_1 - n_2$ units in the original survey who were not included in the remeasurement study. Further, the estimator does not incorporate information on auxiliary variables, x, which may be combined with the information on y and μ available from the survey to provide a more precise estimator of response bias.

Given that we have a stratified, two-phase sample design and resulting data (y, μ, x), our objective is to determine the "best" estimator of measurement bias given these data. Our basic approach is to identify a model for the true value, μ_i, which is a function of the observed values, y_i, $i = 1, \ldots, n_1$, and any auxiliary information, x, that may be available for the population. The model is then used to predict μ_i for all units in the population for which μ_i is unknown. These predictions can then be used to obtain estimates of the true population mean, total, or proportion. Thus, estimators of the response bias for these parameters can be derived from the main survey. Since the approach provides a prediction equation for μ_i which is a function of the observations, estimators of response bias can be computed for areas having small sample sizes. In this case, the prediction equation for μ_i may be augmented by other respondent variables such as demographic characteristics, type of unit, unit size, geographic characteristics, and so on.

The basic estimation and evaluation theory for a prediction approach to the estimation of response bias is presented in the following sections. Under stratified random sampling, estimators of means and totals, their variances and their mean squared errors are provided. Results from application to National Agricultural Statistics Service (NASS) data are also presented.

2. METHODOLOGY FOR ESTIMATION AND EVALUATION

2.1 The Measurement Error Model

To fix the ideas, we shall consider the case of simple random sampling without replacement (SRSWOR) from a single population. Generalizations to stratified random sampling are straightforward and will be considered subsequently.

Let $U = \{1, 2, \ldots, N\}$ denote the label set for the population and let $S_1 = \{1, 2, \ldots, n_1\}$, without loss of generality, denote the label set for the first phase SRSWOR sample of n_1 units from U.

For y_i, $i \epsilon S_1$, assume the model

$$y_i = \gamma_0 + \gamma\mu_i + \epsilon_i, \tag{2.1}$$

where μ_i is the true value of the measured characteristic, γ_0 and γ are constants, and ϵ_i is an independent error term having zero expectation and conditional variance, $\sigma_{\epsilon i}^2$.

Since the focus of our investigation is on the bias associated with the measurements y_i, consider the expectation of y_i. Let $\mathrm{E}(y_i \mid i)$ denote the conditional expectation of y_i over the distribution of the ϵ_i holding the unit i fixed and let $\mathrm{E}(y_i) = \mathrm{E}_i[\mathrm{E}(y_i \mid i)]$ denote the expectation of $\mathrm{E}(y_i \mid i)$ over the sampling distribution. Then, for a given unit, i,

$$\mathrm{E}(y_i \mid i) = \gamma_0 + \gamma\mu_i \tag{2.2}$$

and, hence, the unconditional expectation is

$$\mathrm{E}(y_i) = \gamma_0 + \gamma\bar{M}, \tag{2.3}$$

where $\bar{M} = \sum_{i=1}^{N} \mu_i/N$. Thus, the measurement bias is

$$\bar{B} = \mathrm{E}(y_i - \mu_i) = \gamma_0 + (\gamma - 1)\bar{M}. \tag{2.4}$$

The parameter, γ_0, is a constant bias term that does not depend upon the magnitude of $\bar{M}$. Note that this definition of γ_0 is consistent with the usual definition of measurement bias obtained from the simple model

$$y_i = \mu_i + \epsilon_i, \tag{2.5}$$

with $\epsilon_i \sim (\gamma_0, \sigma_{\epsilon i}^2)$. (See, for example, Biemer and Stokes 1991.)

Consider the estimation of $\bar{B}$. Assume that a subsample of size n_2 of the original n_1 sample units is selected and the true value, μ_i, is measured for these n_2 units. The true value may be ascertained either by a reinterview, a record check, interviewer observation, or some other means. Let $S_2 \subseteq S_1$ denote this so-called second phase sample. The usual estimator of the measurement bias is the NDR defined in (1.1). If the assumption that "the true value, μ_i, is observed in phase 2, for all $i \epsilon S_2$" is satisfied, then the NDR is an unbiased estimator of $\bar{B}$. It may further be shown that the variance of the NDR is

$$\mathrm{E}\left\{\left(1 - \frac{n_2}{n_1}\right)\frac{s_\mu^2}{n_2}\left(1 - \frac{s_{\mu y}^2}{s_y^2 s_\mu^2}\right)\right. \\ \left. + \left(1 - \frac{n_2}{n_1}\right)\frac{s_y^2}{n_2}(1 - r)^2\right\}, \tag{2.6}$$

where $s_\mu^2 = \sum_{j\epsilon S_2} (\mu_j - \bar{\mu}_2)^2/(n_2 - 1)$ with analogous definitions for s_y^2 and $s_{\mu y}$, and $r = s_{\mu y}/s_y^2$.

The NDR may be suboptimal in a number of situations which occur with some frequency. To see this, consider estimators of $\bar{B}$ of the form

$$\bar{b}_{ga} = \bar{y}_g - \bar{\mu}_{Ra}, \tag{2.7}$$

where $\bar{y}_g = \sum_{j\epsilon s_g} y_j/n_g$, $g = 1, 2$,

$$\bar{\mu}_{Ra} = \bar{\mu}_2 + a(\bar{y}_1 - \bar{y}_2) \tag{2.8}$$

and $\bar{\mu}_2 = \sum_{j\epsilon s_2} \mu_j/n_2$, for a a constant given the subsample, S_1. It can be shown that the value of a that minimizes $\mathrm{Var}(\bar{b}_{ga})$ is

$$\begin{aligned} &a = r \qquad \text{for} \quad g = 1, \\ \text{or} \\ &a = r - 1 \quad \text{for} \quad g = 2. \end{aligned} \tag{2.9}$$

Thus, for $g = 1$ or 2, the "optimal" choice of $\bar{b}_{ga}$ is

$$\bar{b}_{\text{opt}} = \bar{y}_1 - [\bar{\mu}_2 + r(\bar{y}_1 - \bar{y}_2)], \tag{2.10}$$

which differs from the NDR by the term $(r-1)(\bar{y}_1 - \bar{y}_2)$. Since, in general, $\bar{y}_1 \neq \bar{y}_2$, NDR is optimal only if $r = 1$. It can be shown that this corresponds to the case where γ_1 in (2.1) is 1.

In this paper we shall explore alternatives to the NDR which incorporate information on y for units in the set $S_1 \sim S_2$ as well as information on some auxilliary variable, x. To illustrate the concepts, we shall restrict ourselves to "no-intercept" linear models initially, *i.e.*, models for which $\gamma_0 = 0$ in (2.1). This important class of models includes the difference estimator as well as ratio estimators.

2.2 Model Prediction Approaches To Estimation

Model prediction approaches to the estimation of population parameters in finite population sampling are well-documented in the literature. Cochran (1977) and other authors have demonstrated the model-based foundations of the ubiquitous ratio estimator. There is also considerable literature on the choice between using weights that are derived from explicit model assumptions in estimation for complex surveys or eliminating the sample weights. Proponents of so-called model-based estimation recommend against the use of weights in parameter estimation (see, for example, Royall and Herson 1973; and Royall and Cumberland 1981). They contend that the probabilities of selection in finite population sampling, whether equal or unequal, are irrelevant once the sample is produced. The reliability criteria used by model-based samples are derived from the model distributional assumptions rather than sampling distributions. If an appropriate model is chosen to describe the relationship between the response variable and other measured survey variables, "model-unbiased" estimators of the population parameters may be obtained which have greater reliability than estimators which incorporate weights.

On the other side of the controversy are the design-based samplers. Instead of the model-based assumptions, design-based samplers assume that an estimator from a survey is a single realization from a large population of potential realizations of the estimator, where each potential realization depends upon the selected sample. The distribution of the values of the estimator when all possible samples that may be selected by the sampling scheme are considered is referred to as the sampling distribution of the estimator. Criteria for evaluating estimators under the design-based approach then consider the properties of the sampling distributions of the estimators. Under this approach, weighting of the estimators is required to achieve unbiasedness if unequal probability sampling is used.

Although the estimators of $\bar{B}$ considered here represent all three classes of estimators, the objective of this paper is not necessarily to compare design-based, model-assisted, and model-based estimators. Rather, we first seek to develop a systematic approach for evaluating alternative estimators for a given two-phase sample design. The problem considered is the following: Given a two-phase sample design and estimators of $B = N\bar{B}$ denoted by $\hat{B}_1$, $\hat{B}_2, \ldots, \hat{B}_p$, how does an analyst identify which estimator minimizes the mean squared error? A second objective of the article is to specify a number of alternative estimators, and apply a systematic approach for evaluating the estimators. As an illustration, the methodology will be applied to data from the National Agricultural Statistics Service's December 1990 Agricultural Survey.

2.3 The Estimators Considered in Our Study

Extending the previously developed notation to stratified, two-phase designs, let N_h denote the size of the hth stratum, for $h = 1, \ldots, L$. A two-phase sample is selected in each stratum using simple random sampling at each phase. Let n_{1h} and $n_{2h} \leq n_{1h}$ denote the phase 1 and phase 2 sample sizes, respectively, in stratum h. Let S_{1h} and $S_{2h} \subseteq S_{1h}$ denote the label sets for the phase 1 and phase 2 samples, respectively, in stratum h. Assume the following data are either observed or otherwise known:

outcome variables:	$y_i \ \forall \ i\epsilon S_{1h}$
true values:	$\mu_i \ \forall \ i\epsilon S_{2h}$
auxilliary variables:	$x_i \ \forall \ i\epsilon S_{1h}$.

Further assume that $X_h = \sum_{i\epsilon U_h} x_i$ is known for $h = 1, \ldots, L$ where U_h is the label set for the hth stratum.

2.3.1 Weighted Estimators of M and B

As a matter of convenience, we shall consider the estimation of the bias for an estimator of a population total denoted by M. The usual estimator of $M = N\bar{M}$ is the unbiased stratified estimator given by

$$\hat{M}_{2st} = \sum_h N_h \bar{\mu}_{2h}, \tag{2.11}$$

where $\bar{\mu}_{2h} = \sum_{i \in S_{2h}} \mu_i / n_{2h}$. The corresponding estimator of $B = N\bar{B}$ is N times the NDR defined in (1.1). For stratified samples, it is

$$\hat{B}_{2st} = \hat{Y}_{2st} - \hat{M}_{2st}, \tag{2.12}$$

where $\hat{Y}_{2st} = \sum_h N_h \bar{y}_{2h}$ and $\bar{y}_{2h} = \sum_{i \in S_{2h}} y_i / n_{2h}$. Note that (2.12) does not incorporate the information on y for units with labels $i \in S_{1h} \sim S_{2h}$. An alternative estimator that uses all the data on y is

$$\hat{B}_{12st} = \hat{Y}_{1st} - \hat{M}_{2st}, \tag{2.13}$$

where $\hat{Y}_{1st} = \sum_h N_h \bar{y}_{1h}$ and $\bar{y}_{1h} = \sum_{i \in S_{1h}} y_i / n_{1h}$.

A number of model-assisted estimators can be specified for two-phase stratified designs. These may take the form of either separate or combined estimators (see, for example, Cochran 1977, pp. 327-330). Further, the ratio adjustments may be applied to either phase 1 or phase 2 stratum-level estimators. Because stratum sample sizes are typically small in two-phase samples, only combined estimators shall be considered here.

As the emphasis in this paper is on the development of the methodology for model-based estimates of measurement bias and their evaluation, we shall consider a simple, special case of the model (2.1); *viz.*, $\gamma_0 = 0$ or the no-intercept model. However, generalizations of the no-intercept methodology to multivariate intercept models do not afford any difficulties and will be considered in a subsequent paper. Thus, letting $\gamma_0 = 0$ in (2.1) we have

$$y_i = \gamma \mu_i + \epsilon_i, \tag{2.14}$$

where γ is an unknown constant and we assume $\epsilon_i \sim (0, \sigma_\epsilon^2 \mu_i)$. The least squares estimator of γ is $\hat{\gamma} = \bar{y}_{2st}/\bar{\mu}_{2st}$, where $\bar{y}_{2st} = \hat{Y}_{2st}/N$ and $\bar{\mu}_{2st} = \hat{M}_{2st}/N$. Thus, a model-assisted estimator of μ_i is $y_i/\hat{\gamma} = \bar{\mu}_{2st} y_i / \bar{y}_{2st}$ and of M is

$$\hat{M}_{2stR} = \frac{\hat{M}_{2st}}{\hat{Y}_{2st}} \hat{Y}_{1st}. \tag{2.15}$$

Using this estimator of M, two estimators of B corresponding to (2.12) and (2.13) are

$$\hat{B}_{2stR} = \hat{Y}_{2st} - \hat{M}_{2stR} \tag{2.16}$$

and

$$\hat{B}_{12stR} = \hat{Y}_{1st} - \hat{M}_{2stR}. \tag{2.17}$$

A third estimator of B can be obtained via the model

$$y_i = \beta x_i + e_i, \tag{2.18}$$

where β is a constant and $e_i \sim (0, \sigma_e^2 x_i)$. This leads to a ratio estimator of Y,

$$\hat{Y}_{xstR} = \frac{\bar{y}_{1st}}{\bar{x}_{1st}} X. \tag{2.19}$$

Thus, the corresponding estimator of B is

$$\hat{B}_{x2stR} = \hat{Y}_{xstR} - \hat{M}_{2stR}. \tag{2.20}$$

Finally, Särndal, Swensson and Wretman (1992, p. 360) suggest a general estimator of M in two-phase sampling. Applying their equation 9.7.2 to the model in (2.14) under stratified sampling yields

$$\hat{M}_{SSW} = \hat{M}_{2stR} + \frac{\bar{\mu}_{2st}}{\bar{x}_{2st}} (X - \hat{X}_{1st}). \tag{2.21}$$

Note that this estimator is simply (2.15) with the addition of the unbiased estimator of zero. The resulting estimator may have smaller variance than $\hat{M}_{2stR}$ if this term is negatively correlated with $\hat{M}_{2stR}$. Likewise, their estimator of Y reduces to $\hat{Y}_{xstR}$ defined in (2.19). Thus the corresponding estimator of B is

$$\hat{B}_{SSW} = \hat{Y}_{xstR} - \hat{M}_{SSW}, \tag{2.22}$$

which is identical to $\hat{B}_{SSW} = B_{x2stR}$ plus the second term of the right hand side of (2.21).

2.3.2 Unweighted Estimators of M and B

Rewrite M as

$$\begin{aligned} M &= \sum_{i \in S_2} \mu_i + \sum_{i \in S_1 \sim S_2} \mu_i + \sum_{i \in U \sim S_1} \mu_i \\ &= M_{(2)} + M_{(1 \sim 2)} + M_{(\sim 1)}, \end{aligned} \tag{2.23}$$

say, where $S_g = \cup_{h=1}^{L} S_{gh}$, $g = 1, 2$. The strategy for unweighted, model-based estimation is to replace μ_i in $M_{(1 \sim 2)}$ and $M_{(\sim 1)}$ by a prediction, $\hat{\mu}_i$, obtained from a model.

Using the model in (2.14), an estimator of μ_i is

$$\hat{\mu}_i = y_i / \bar{\gamma},$$

where now $\bar{\gamma} = \bar{y}_2 / \bar{\mu}_2$. Thus, an estimator of $M_{(1 \sim 2)}$ is

$$\hat{M}_{(1\sim 2)} = \frac{\bar{\mu}_2}{\bar{y}_2} \sum_{i \in S_1 \sim S_2} y_i = \frac{\bar{\mu}_2}{\bar{y}_2} (n_1 \bar{y}_1 - n_2 \bar{y}_2), \tag{2.24}$$

where $\bar{y}_g = \sum_{i \in S_g} y_i / n_g$, $\bar{\mu}_2 = \sum_{i \in S_2} \mu_i / n_2$, and $n_g = \sum_h n_{gh}$, for $g = 1, 2$. Further, using the model

$$\mu_i = \delta x_i + \xi_i, \tag{2.25}$$

where δ is a constant and $\xi_i \sim (0, \sigma_\xi^2 x_i)$, we obtain

$$\hat{M}_{(\sim 1)} = \frac{\bar{\mu}_2}{\bar{x}_2} X_{U \sim S_1}, \tag{2.26}$$

where $X_{U \sim S_1} = \sum_{i \in U \sim S_1} X_i$. Thus, a model based estimator of M is

$$\hat{M}_M = M_{(2)} + \hat{M}_{(1 \sim 2)} + \hat{M}_{(\sim 1)} = \hat{M}_{(1)} + \hat{M}_{(\sim 1)}, \tag{2.27}$$

where $\hat{M}_{(1)} = n_1 \bar{\mu}_2 \bar{y}_1 / \bar{y}_2$.

Likewise, Y can be rewritten as

$$Y = \sum_{i \in S_1} y_i + \sum_{i \in U \sim S_1} y_i = Y_{(1)} + Y_{(\sim 1)} \tag{2.28}$$

and we wish to predict y_i in $Y_{(\sim 1)}$. Using the model in (2.18) a model-based estimator of $Y_{(\sim 1)}$ is

$$\hat{Y}_{(\sim 1)} = \frac{\bar{y}_1}{\bar{x}_1} X_{U \sim S_1}$$

and, thus, an estimator of Y is

$$\hat{Y}_M = Y_{(1)} + \hat{Y}_{(\sim 1)}. \tag{2.29}$$

Thus, B is estimated as

$$\hat{B}_M = \hat{Y}_M - \hat{M}_M. \tag{2.30}$$

Versions of $\hat{B}_{2stR}$, $\hat{B}_{12stR}$, $\hat{B}_{x2stR}$ and $\hat{B}_M$ which are more robust to model outliers may also be constructed. The corresponding estimators, denoted by $\tilde{B}_{2stR}$, $\tilde{B}_{12stR}$, $\tilde{B}_{x2stR}$ and $\tilde{B}_M$, respectively, may be formed by eliminating those data points which deviate substantially from the model predictions and computing the model-based or model-assisted estimators using the remaining data. To illustrate, consider the estimator $\hat{M}_{2stR}$ in (2.15). For this estimator, let

$$(n_{2h} - 1) s_{res,h}^2 = \sum_{\mu_{hi} \neq 0} \frac{(y_{hi} - \hat{\gamma}\, \mu_{hi})^2}{\mu_{hi}}, \tag{2.31}$$

denote the sum of squares of residuals for the model (2.14). Then, in calculating the estimator of γ, only those units in $i \in \tilde{S}_{2h}$ where $\tilde{S}_{2h} = \{ i \in S_{2h} : | y_{ih} - \hat{\gamma} \mu_{ih} | \leq 3 s_{res,h} \sqrt{\mu_{hi}} \}$ are used. Denoting this estimator of γ as $\tilde{\gamma}$, the estimator of M is $\tilde{M}_{2stR} = \hat{Y}_{1st} / \tilde{\gamma}$ where $\tilde{\gamma} = \tilde{y}_{2st} / \tilde{\mu}_{2st}$ and $\tilde{\mu}_{2st}$ and $\tilde{y}_{2st}$ are the stratified means of μ_i and y_i for $i \in \tilde{S}_{2h}$. The other robust model prediction estimators may be computed analogously.

Many other unweighted, model-based estimators may be explored in the context of our two-phase design. For example, an intercept term may be added to models (2.14), (2.18), and (2.25). Further, slope and intercept parameters may be specified separately for each stratum or combination of strata.

2.4 Estimation of Mean Squared Errors Using Bootstrap Estimators

Although it is possible, under the appropriate design-based or model-based assumptions, to derive closed form analytical estimates of the variance of the estimators we are considering in this study, we have elected instead to use a computer-intensive resampling method. First, we seek a method which is easy to apply since there are potentially many estimators which will be considered in our study. Secondly, it is important to evaluate each estimator using the same criteria and a consistent method of variance estimation is essential to achieving this objective. Thus, it is essential that we employ a variance estimation method which can be applied to estimators of any complexity, under assumptions which are consistent and which do not rely upon any model assumptions. It is well-known that model-based variance estimation approaches are quite sensitive to model failure (see, for example, Royall and Herson 1973; Royall and Cumberland 1978; and Hansen, Madow and Tepping 1983). Royall and Cumberland (1981) discuss several bias relevant alternatives including the jackknife variance estimator.

Our approach is similar to that of Royall and Cumberland except rather than using a jackknife estimator, we employ a bootstrap estimator of the variance. For independent and identically distributed observations, Efron and Gong (1983) show that the bootstrap and the jackknife variance estimators differ by a factor of $n/(n - 1)$ for samples of size n. Thus, the robustness properties Royall and Cumberland demonstrate for the jackknife estimator also hold for the bootstrap estimator.

Other properties of the bootstrap estimator have led us to choose it above other resampling methods. The jackknife and balance repeated replication (BRR) methods are not easily modified for the two-phase sampling design of our study. However, the bootstrap is readily adaptable to two-phase sampling. Further, Rao and Wu (1988) provide evidence from a simulation study that the coverage properties of bootstrap confidence intervals in complex sampling compare favorably to the jackknife and BRR.

Our general approach extends the method developed by Bickel and Freedman (1984) for single phase, stratified sampling, to two-phase stratified sampling. Since the bootstrap procedure is implemented independently for each stratum, we shall, for simplicity, describe the method for the single stratum case.

2.4.1 Estimation of Variance

Extending the bootstrap method to two-phase sampling is not simply a matter of subsampling the single phase bootstrap samples. Recall that true values are known only for the units in S_2 and, therefore, the bootstrap sampling scheme must necessarily confine the selection to units in S_2. Therefore, let S_1 and S_2 denote the phase 1 and phase 2 samples, respectively, selected from U using SRSWOR. Let $S_{1\sim2}$ denote the label set, $S_1 \sim S_2$. Let $\hat{\Theta} = \hat{\Theta}(S_{1\sim2}, S_2)$ denote an estimator of Θ which may be a function of the observations corresponding to units in both S_2 and $S_{1\sim2}$. Define N, n_1, n_2 and $n_{1\sim2}$ as the sizes of sets U, S_1, S_2 and $S_{1\sim2}$, respectively. Consider how the bootstrap is applied to obtain estimates of $\mathrm{Var}(\hat{\Theta})$.

The simplest case is when N/n_1 is an integer, say k. First, we form the psuedo-population label set

$$U_A^* = U_{A(2)}^* \cup U_{A(1\sim2)}^*, \tag{2.32}$$

where $U_{A(2)}^*$ consists of k copies of the units in S_2 and $U_{A(1\sim2)}^*$ consists of k copies of the units in $S_{1\sim2}$. We then perform the following three steps:

1. Draw a SRSWOR of size n_2 from $U_{A(2)}^*$ and denote this set by S_2^*.
2. Draw a SRSWOR of size $n_{1\sim2}$ from $U_{A(1\sim2)}^*$ and denote this set by $S_{1\sim2}^*$.
3. Compute $\hat{\Theta}_1^* = \hat{\Theta}_1(S_{1\sim2}^*, S_2^*)$ which has the same functional form as $\hat{\Theta}(S_{1\sim2}, S_2)$, but is computed for the $n_1 = n_{1\sim2} + n_2$ units in $S_1^* = S_{1\sim2}^* \cup S_2^*$.

Repeat steps 1 to 3 some large number, Q, times to obtain $\Theta_1^*, \ldots, \Theta_Q^*$. Then, an estimator of $\mathrm{Var}(\hat{\Theta})$ is

$$\mathrm{var}_{BSS}(\hat{\Theta}) = \sum_{q=1}^{Q} \frac{(\hat{\Theta}_q^* - \hat{\Theta}_{\cdot}^*)^2}{Q - 1}, \tag{2.33}$$

where $\hat{\Theta}_{\cdot}^* = \sum_{q=1}^{Q} \hat{\Theta}_q^*/Q$.

Using the methods of Rao and Wu (1988), it can now be shown that $\mathrm{var}_{BSS}(\hat{\Theta})$ is a consistent estimator of $\mathrm{Var}(\hat{\Theta})$. If $N = kn_1 + r$, where $0 < r < n_1$, the procedure is modified as follows using the Bickel and Freedman procedure. First, form the pseudo-population U_A^* as above consisting of kn_1 units. In addition, form the pseudo population $U_B^* = U_{B(1\sim2)} \cup U_{B(2)}^*$ of size $(k + 1)n_1$ where $U_{B(1\sim2)}^*$ and $U_{B(2)}$ consist of $k + 1$ copies of the labels in $S_{1\sim2}$ and S_2, respectively. Then, for αQ of the bootstrap samples, select $S_1^* = S_{1\sim2}^* \cup S_2^*$ from U_A^* and for $(1 - \alpha)Q$ samples, select S_1^* from the psuedo-population, U_B^* using the three-step procedure described above, where

$$\alpha = \left(1 - \frac{r}{n_1}\right)\left(1 - \frac{r}{N - 1}\right). \tag{2.34}$$

2.4.2 Estimation of Bias and MSE

The bootstrap procedure can also provide an estimate of estimator bias. The usual bootstrap bias estimator (see Efron and Gong 1983; Rao and Wu 1988) is $b(\hat{\Theta}) = \hat{\Theta}_{\cdot}^* - \hat{\Theta}$ where $\hat{\Theta}_{\cdot}^* = \sum_q \hat{\Theta}_q^*/Q$ and $\hat{\Theta}$ is the estimate computed from the full sample. Note that $\hat{\Theta}_q^*(q = 1, \ldots, Q)$ and $\hat{\Theta}$ have the same functional form and are based upon the same model assumptions. Thus $b(\hat{\Theta})$ does not reflect the contribution to bias due to model failure. We propose an alternative estimator of bias which we conjecture is an improvement over $b(\hat{\Theta})$.

Recall from (2.4) that $\bar{B} = \mathrm{E}(y_i - \mu_i)$ where $\mathrm{E}(\cdot)$ denotes expectation over both the measurement error and sampling error distributions. Thus, $\bar{B}$ may be rewritten as $\bar{B} = \sum_{i=1}^{N} (Y_i - \mu_i)/N$ where $Y_i = \mathrm{E}(y_i \mid i)$. Since Y_i is unknown and unobservable for all $i \in U$, $\bar{B}$ is also unknown and unobservable. Therefore, we shall construct a pseudo population resembling U, denoted by U^*, such that $\bar{B}^* = \mathrm{E}^*(y_i - \mu_i)$ is known, where $\mathrm{E}^*(\cdot)$ is expected value with respect to both the measurement error and the sampling distributions associated with U^*.

Let $U^* = \cup_{h=1}^{L} U_h^*$ where U_h^* consists of $k_h = N_h/n_{1h}$ copies of the units in S_{1h}. Here we have assumed k_h is an integer, but we will subsequently relax the assumption. Further, denote by y_i^* the value of the characteristic for the unit $i \in U^*$. This value is equal to the y_i for the corresponding unit in S_1. Thus, the population total of the y_i^* is $Y^* = \sum_{i \in U^*} y_i^* = \hat{Y}_{1st}$ for $\hat{Y}_{1st}$ defined in (2.13). Analogously, define the true value for unit $i \in U^*$ as $\mu_i^* = \mu_j$ for $i \in U^*$ corresponding to $j \in S_2$. For $j \in S_{1\sim2}$, μ_j is unknown; however, for our pseudo-population we could generate pseudo-values for the μ_i^* such that $M^* = \sum_{i \in U^*} \mu_i^* = \hat{M}_{2st}$ where $\hat{M}_{2st}$ is defined in (2.11). Thus, for U^*, $B^* = \hat{Y}_{1st} - \hat{M}_{2st} = \hat{B}_{12st}$ defined in (2.13). As we shall see, it is not necessary to generate the pseudo-values for μ_i^* in order to evaluate the bias in the estimators of B^*.

Note that under stratified sampling, $U^* = U_A^*$, as defined in Section 2.4. Further, the bootstrap procedure described in this section is equivalent to repeated sampling from U^* and the alternative estimators $\hat{\Theta}_1, \ldots, \hat{\Theta}_p$ of B

may also be considered estimators of B^*. Since B^* is known, the bias of $\hat{\theta}$ as an estimator of B^* is $\hat{B}^* = \hat{\theta} - B^*$ and the corresponding MSE may be estimated as

$$\widehat{\text{MSE}} = \sum_q (\hat{\theta}_q - B^*)^2/Q$$

$$\doteq \text{var}_{BSS}(\hat{\theta}) + (\hat{\theta}^*_{\cdot} - B^*)^2, \qquad (2.35)$$

where $\text{var}_{BSS}(\hat{\theta})$, $\hat{\theta}_q$, and $\hat{\theta}^*_{\cdot}$ are defined in Section 2.4. It can be easily verified that these results still hold when k_h is non-integer.

Thus, the bootstrap procedure provides a method for evaluating the MSE of alternative estimators for estimating B^*. Further, the pseudo-population U^* is a reconstruction of U based upon copies of the values for the units in S_1 and S_2. Thus, it is reasonable to use $\hat{B}^*$ and $\widehat{\text{MSE}}^*$ to evaluate alternative estimators of B.

3. APPLICATION TO THE AGRICULTURAL SURVEY

3.1 Description of the Survey

The National Agricultural Statistics Service (NASS) annually conducts a series of surveys which are collectively referred to as the Agricultural Survey (AS) program. The purpose of these surveys is to collect data related to specific agricultural commodities at the state and national levels. Each December in the years 1988-1990, reinterview studies designed to assess the measurement bias in the data collected by Computer Assisted Telephone Interviewing (CATI) were conducted in six states: Indiana, Iowa, Minnesota, Nebraska, Ohio, and Pennsylvania. The reinterview techniques employed in these three studies are very similar to those used by the U.S. Census Bureau (see, for example, Forsman and Schreiner 1991). However, unlike the Census Bureau's program, the major objective in the NASS studies is the estimation of measurement bias rather than interviewer performance evaluation.

As noted above, only AS responding units whose original interview was conducted by CATI were eligible for selection into the reinterview sample. The reasons for this restriction on sampling were primarily cost, timing, and convenience. However, a large proportion of the AS is conducted by CATI and, thus, information regarding AS measurement bias for this group would provide important information for the entire AS program.

For the NASS reinterview studies, the interviewing staff consisted of a mix of field supervisors and experienced field interviewers. This interviewing staff, which was a separate corps of interviewers from those used for CATI, conducted face-to-face reinterviews in a subsample of AS units for a subset of AS survey items. To minimize any problems that respondents may have with recall, the reinterviews were conducted within 10 days of the original interview. Differences between the original AS and reinterview responses were reconciled to determine the "true" value. Considerable effort was expended in procedural development, training, and supervision of the reinterview process to ensure that the final reconciled response was completely accurate. For the most part, the wording of the subset of AS questions asked in the reinterview was identical to that of the parent survey. The reinterviewers attempted to contact the most knowledgeable respondent in order to ensure the accuracy of the reconciled values.

In this report, only the 1990 data are analyzed. Table 1 presents the reinterview sample sizes for this study.

Table 1

Sample Sizes by Survey Item

Item	x	y	μ
	U	S_1	S_2
All wheat stocks	108,267	8,176	1,157
Corn planted acres	225,269	8,211	1,157
Corn stocks	225,269	7,990	1,115
Cropland acreage	278,045	8,274	1,141
Grain storage capacity	207,460	8,126	1,104
Soybean planted acreage	171,761	8,211	1,156
Soybean stocks	171,761	8,113	1,130
Total land in farm	276,450	8,309	1,159
Total hog/pig inventory	248,571	8,247	1,142
Winter wheat seedings	108,267	8,211	1,150

3.2 Comparison of the Estimators of M and B

Using the December 1990 Agricultural Survey and its corresponding reinterview survey data, the estimators developed in the previous section were compared. Estimates of standard errors and mean squared errors were computed using the Bickel-Freedman bootstrap procedure described in Section 2.4, with $Q = 300$ bootstrap samples. Table 2 displays the results for six of the estimators: $\hat{B}_{2st}$, the traditional difference estimator; $\hat{B}_{x2stR}$, the weighted ratio estimator; $\tilde{B}_{x2stR}$, the robust (outlier deletion) version of $\hat{B}_{x2stR}$; $\hat{B}_{SSW}$, the Särndal, Swensson and Wretman estimator; $\hat{B}_M$, the unweighted model-based estimator; and $\tilde{B}_M$, the robust (outlier deletion) version of $\hat{B}_M$.

3.3 Summary of Results

Table 2 presents a summary of the results from our study. The first data column is the known value of $B^* = \mathrm{E}(y_i^* - \mu_i^*)$, the bias parameter for the pseudo-population, U^*. The other data columns contain the values of the estimators with their standard errors in parentheses, where s.e. $(\hat{\Theta}) = \sqrt{\mathrm{var}_{BSS}(\hat{\Theta})}$. The last four rows of the table correspond, respectively, to:

(a) the number of items (out of 10) for which a 95% confidence interval contains B^*;
(b) the average coefficient of variation (C.V.);
(c) the average square root of $\widehat{\mathrm{MSE}}$ (RMSE); and
(d) the average absolute relative bias.

A striking feature of these results is the large disparity among the six estimators across all commodities; particularly for All Wheat Stocks. For this commodity, the range of estimates is −94.2 to 103.2. Also indicated (by the ‡ symbol) in Table 2 is whether a 95% confidence interval, *i.e.*, $[\hat{\Theta} - 2\,\mathrm{s.e.}\,(\hat{\Theta}), \hat{\Theta} + 2\,\mathrm{s.e.}\,(\hat{\Theta})]$, covers the parameter B^*. The best performer for parameter coverage is $\hat{B}_{SSW}$ which produced confidence intervals that covered B^* for eight out of ten commodities. $\hat{B}_{2st}$ was the next best with six and $\hat{B}_M$ was third with five. The traditional ratio estimator and its robust version were the worst performers with only one commodity having a confidence interval covering B^*.

Application of the mean squared error criterion presents a different picture. Here, $\tilde{B}_M$ emerged as the estimator having the smallest average root MSE. However, $\hat{B}_{SSW}$ and $\hat{B}_{2st}$ are not much greater. Further, $\hat{B}_{SSW}$ was the estimator having the smallest average absolute relative bias. Only two commodities were estimated with significant biases using this estimator. Thus, it appears from these results that $\hat{B}_{SSW}$ is the preferred estimator using overall performance as the evaluation criterion.

Table 2

Comparison of Estimators with, B^*, the Pseudo-Population Value of the Bias†

Characteristic	B^*	$\hat{B}_{2st}$	$\hat{B}_{x2stR}$	$\tilde{B}_{x2stR}$	$\hat{B}_{SSW}$	$\hat{B}_M$	$\tilde{B}_M$
All wheat stocks	42.3	−6.1 (12.3)	103.2 (17.6)	−94.2 (16.5)	−0.9‡ (24.8)	19.2‡ (16.5)	10.6‡ (16.7)
Corn planted acreage	−1.8	1.1‡ (1.1)	11.7 (1.3)	10.1 (1.1)	0.3‡ (1.2)	−4.7‡ (1.9)	−5.0 (1.5)
Corn stocks	−6.4	−5.4‡ (1.5)	2.4 (1.6)	0.2 (1.3)	−6.5‡ (1.6)	−7.9‡ (2.4)	−9.3‡ (2.2)
Cropland acreage	27.0	−19.6 (8.3)	−15.0 (8.3)	7.0 (3.1)	−19.6 (8.2)	−36.8 (11.0)	−12.8 (4.0)
Grain storage capacity	−3.37	1.4‡ (3.7)	32.3 (3.7)	29.5 (2.6)	−0.1‡ (3.9)	−6.9 (3.0)	−6.8 (2.5)
Soybean planted acreage	−4.4	0.8 (0.8)	13.0 (1.0)	9.9 (0.9)	−0.3 (1.0)	−2.9 (1.1)	−2.7 (1.0)
Soybean stocks	−0.01	2.8‡ (3.1)	21.3 (2.9)	5.0 (2.3)	0.2‡ (3.5)	−11.0 (3.6)	−8.8 (3.4)
Total land in farm	−20.0	−24.7‡ (10.4)	−18.8‡ (12.5)	−2.6 (7.6)	−25.7‡ (10.7)	−44.5‡ (13.4)	−21.2 (5.8)
Total hogs/pigs inventory	−0.1	−2.1 (0.9)	3.4 (1.1)	−0.0‡ (1.0)	−2.2‡ (1.1)	−2.5‡ (1.3)	−1.6‡ (1.0)
Winter wheat seedings	−0.6	−0.5‡ (0.4)	3.8 (0.6)	1.8 (0.5)	−1.2‡ (0.6)	1.1 (0.4)	1.1 (0.4)
Number of items where C.I. covers B^*		6	1	1	8	5	3
Average C.V.		1.01	.30	11.1	9.5	.41	.48
Average RMSE		13.2	22.4	25.2	12.9	14.9	10.8
Average | Relbias |		30.8	220.0	53.4	4.9	113.1	91.3

† Standard errors in parentheses.
‡ 95% confidence interval covers the pseudo population parameter.

4. CONCLUSIONS AND RECOMMENDATIONS

In this article, we developed a general methodology for constructing and evaluating weighted and unweighted model prediction estimators of measurement bias for stratified random, two-phase sample designs. The proposed estimators incorporate information on the observations, y, from the first phase sample, and an auxilliary variable, x. Model robust versions of the estimators were also considered and evaluated. The ultimate goal of model prediction estimation is to identify estimators which make "optimal" use of the data (y, μ, x). The general estimation and evaluation methodology for achieving this goal was illustrated for the ordinary regression model with no intercept. However, the methodology can be easily extended to multivariate, intercept models.

Our proposed evaluation criteria are based upon estimates of bias, variance, and mean squared error computed using a bootstrap resampling methodology. The method of Bickel and Freedman was extended to two-phase sampling for this purpose. It was shown both analytically and empirically that the usual NDR estimator is not optimal under the model prediction approach to estimating measurement bias. Our analyses found that, for the six estimators we considered, the estimator derived from the work of Särndal *et al.* (1992), was the best overall estimator by the bootstap evaluation criteria.

Incorporating auxiliary information into the estimation of measurement bias creates a number of practical problems which may increase the costs and reduce the timeliness of producing the estimates. First, the auxiliary variable, x, must be available, at least in aggregate form, for all socioeconomic and geographic domains for which model prediction estimates are desired. This could be a large data management task. Further, the complexity of the variance estimator using analytical methods increases with the complexity of the bias estimator. Although simpler, the bootstrap variance estimation method can be prohibitively expensive if computer time must be purchased. However, these difficulties are not insurmountable, especially if a high-powered microcomputer is available. Further, given the cost of reinterview surveys for estimating measurement bias, even moderate increases in precision in the bias estimators can result in substantial cost savings.

The model prediction approach has the potential for extracting the maximum information on response bias from reinterview surveys and thus model prediction estimators will usually be more efficient than the traditional net difference estimator. In addition, the model prediction approach may also offer a means for extrapolating estimates of bias to areas which were not sampled. As an example, in the NASS application, the reinterview sample was drawn only from the CATI areas for reasons of operational convenience and cost efficiency. However, by using prediction models which are functions of the original responses and other available characteristics, it may be possible to predict the measurement bias in the non-CATI survey areas from the local characteristics of these areas - a type of "synthetic" estimation. Although this application of model-based estimation was not considered in this paper, it is a natural extension of the methodology and one which will be evaluated in a subsequent study.

Also for future research, we intend to incorporate multivariate, intercept models in the estimation of measurement bias. Since the bootstrap evaluation criteria developed in this article are general, no changes in the evaluation methodology are required to handle the addition of variables in the estimation models. Further, the model assumptions and the methods for handling outliers will be refined and evaluated in a subsequent paper. Finally, we need to explore the effect on estimation of departures from the model assumptions, particularly the assumption that the reinterview observation is without error. As Fuller (1991) has shown, if the reinterview is fallible but unbiased, the variance of the predicted values increases but the predictions are still unbiased. Thus, under these assumptions, one could explore the relative precision of the alternative estimators of measurement bias in order to determine the robustness of the model prediction approach.

ACKNOWLEDGEMENTS

The authors thank Manuel Cárdenas for his valuable assistance and the referees for their helpful comments.

REFERENCES

BICKEL, P., and FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

BIEMER, P., and STOKES, L. (1991). Approaches to the modeling of measurement errors. In *Measurement Errors in Surveys*, (Eds. P. Biemer, *et al.*). New York: John Wiley and Sons.

COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

EFRON, B., and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 31, 36-48.

FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*, (Eds. P. Biemer, *et al.*). New York: John Wiley and Sons.

FULLER, W.A. (1991). Regression estimation in the presence of measurement error. In *Measurement Errors in Surveys*, (Eds. P.P. Biemer, *et al.)*. New York: John Wiley and Sons, 617-636.

HANSEN, M., MADOW W., and TEPPING, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

RAO, J.N.K., and WU, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

ROYALL, R., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-893.

ROYALL, R., and CUMBERLAND, W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-361.

ROYALL, R., and CUMBERLAND, W. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 137–147
Statistics Canada

Model-Based Estimation of Record Linkage Error Rates

J.B. ARMSTRONG and J.E. MAYDA[1]

ABSTRACT

Record linkage is the matching of records containing data on individuals, businesses or dwellings when a unique identifier is not available. Methods used in practice involve classification of record pairs as links and non-links using an automated procedure based on the theoretical framework introduced by Fellegi and Sunter (1969). The estimation of classification error rates is an important issue. Fellegi and Sunter provide a method for calculation of classification error rate estimates as a direct by-product of linkage. These model-based estimates are easier to produce than the estimates based on manual matching of samples that are typically used in practice. Properties of model-based classification error rate estimates obtained using three estimators of model parameters are compared.

KEY WORDS: Mixture model; Latent variable model; Iterative scaling.

1. INTRODUCTION

Computer files containing information about individuals, businesses or dwellings are used in many statistical applications. The linking of records that refer to the same entity is often required. The process of linking records referring to the same entity is called exact matching. If all records involved in an application have been accurately assigned a unique identifier, exact matching is trivial. Record linkage methods deal with the problem of exact matching when a unique identifier is not available. In that case, each record typically includes a number of data fields containing identifying information that could be used for matching. Problems in matching are due to errors in these data or due to the same value for a particular field being valid for more than one entity.

Applications of record linkage include the unduplication of lists of dwellings or businesses obtained from various sources to create survey frames. In addition, record linkage is widely used in applications related to health and epidemiology. Work in this area typically involves matching records containing information on individuals in industrial or occupational cohorts to records documenting the illness or death of individuals. For example, record linkage methodology for follow-up studies of persons exposed to radiation is discussed in Fair, Newcombe and Lalonde (1988).

The record linkage problem can be formulated using two data files that correspond to two populations. Each file may contain information for all entities in the corresponding population or information for a random sample of entities. The file A contains N_A records and the file B contains N_B records. The set of record pairs formed as the cross-product of A and B is denoted by $C = \{(a,b); a \in A, b \in B\}$. C contains $N = N_A \cdot N_B$ record pairs. The objective of record linkage is to partition the set C into two disjoint sets – the set of true matches, denoted by M, and the set of true non-matches, U.

The theoretical framework introduced by Fellegi and Sunter (1969) is the basis of a great deal of applied work. For each record pair, a decision is taken concerning whether or not the records refer to the same entity after examining data recorded on files A and B. The possible decisions are link (A_1), non-link (A_3) and possible link (A_2). There are two types of errors. First, decision A_1 may be taken for a record pair that is a member of U, the set of true non-matches. Second, decision A_3 may be taken for a record pair that is a member of set M, the set of true matches. Acceptable levels of classification error are specified before the files are linked. A record pair is classified as a possible link if the data do not provide sufficient evidence to justify classification of the pair as a link or non-link at error levels less than or equal to those specified. Accurate estimation of classification error rates associated with various decision rules is necessary to determine an appropriate rule. The classification error rate for true non-matches is $P(A_1 \mid U)$. The error rate for true matches is $P(A_3 \mid M)$.

Estimates of classification error rates can be obtained by selecting a sample of record pairs from the set C and manually determining the true match status of sampled pairs. Applications of this approach are described in Bartlett *et al.* (1993). Sampling may be both costly and cumbersome to implement, particularly when the same linkage must be done for a number of pairs of files, each with slightly different characteristics. Belin and Rubin (1991) describe another method of error rate estimation

[1] J.B. Armstrong and J.E. Mayda, Statistics Canada, Business Survey Methods Division, 11-RH Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

that requires true match status for record pairs in a pilot study. In contrast to the straightforward sampling approach, the Belin-Rubin method provides a framework for the application of information obtained from the pilot study to larger linkages involving similar data.

The Fellegi-Sunter framework provides a method for calculation of error rate estimates using estimates of probabilities that record pairs will agree on various combinations of data fields. Calculation of these model-based error rate estimates is straightforward and manual determination of the true match status of record pairs is not required. However, they often have poor properties in applied work. See, for example, Belin (1990). In this paper, the potential for improvement of the properties of model-based error rate estimates through careful estimation of agreement probabilities is examined.

Three alternative estimation methods are evaluated. The approaches described use only the information on files A and B. They do not rely on auxiliary information. Model-based error rate estimates obtained using each alternative are compared with actual error rates using both synthetic data that incorporate important characteristics of data from health applications of record linkage, and information from an actual record linkage application.

The plan of the paper is as follows. Section 2 includes details of the model-based classification error rate estimation method introduced by Fellegi and Sunter. The model for agreement probabilities that forms the basis of subsequent discussion of estimation methods is also specified. Two estimation methods that rely on an important independence assumption are described in Section 3. A third alternative that does not require independence is discussed in Section 4. The results of comparisons of the three approaches using synthetic data are reported in Section 5. The results of evaluation work with information from a real application are described in Section 6. Section 7 contains some concluding remarks.

2. THEORETICAL CONCEPTS

Relevant aspects of the theory for record linkage developed by Fellegi and Sunter (1969) are summarized in this section. In the Fellegi-Sunter framework, estimates of classification error rates are calculated using estimates of probabilities of agreement on various combinations of data fields. Applications of the theory of Fellegi and Sunter usually involve the assumption that the probability that a record pair will agree on a particular data field is independent of the results of comparisons for other fields. The theory is nevertheless very flexible, allowing for any pattern of dependence between results of comparisons for different data fields. A parameterization of dependence in terms of loglinear effects is given.

2.1 Model-Based Classification Error Rate Estimation

To obtain information related to the classification of a record pair as a link (A_1), non-link (A_3) or possible link (A_2), data fields containing identifying information are compared. In an application involving records referring to persons, separate comparisons of family names, given names, and dates of birth might be performed. The outcome of a comparison is a numerical code representing a statement like "names agree", "names disagree", "name missing on one or both files", "names agree and both are George" or "names disagree but their first two characters agree". The outcome codes used in applied work differ between applications and between comparisons in the same application. The smallest number of outcome codes that can be used for any comparison is two - corresponding to agreement and disagreement. An outcome code corresponding to "missing on one or both files" is usually needed in applied work. The agreement outcome may be replaced by a number of value-specific outcomes (such as "names agree and both are George"). Certain disagreements may be coded as partial agreements (such as "names disagree but their first two characters agree").

For present purposes, we consider agreement and disagreement outcomes only. In the case of K matching fields, we introduce the outcome vector $\underset{\sim}{x}^j = (x_1^j, x_2^j, \ldots, x_K^j)$ for record pair j. We have $x_k^j = 1$ if record pair j agrees on data field k and $x_k^j = 0$ if record pair j disagrees on data field k.

Newcombe *et al.* (1959) introduced the idea that decisions concerning whether or not a pair of records represent the same entity should be based on the ratio

$$R(\underset{\sim}{x}) = P(\underset{\sim}{x} \mid M)/P(\underset{\sim}{x} \mid U), \qquad (1)$$

where $\underset{\sim}{x} = (x_1, x_2, \ldots, x_K)$ is the generic outcome vector, $P(\underset{\sim}{x} \mid M)$ is the probability that comparisons for a record pair that is a true match will produce outcome vector $\underset{\sim}{x}$, and $P(\underset{\sim}{x} \mid U)$ is the probability of $\underset{\sim}{x}$ for a record pair that is a true non-match. The optimality of record linkage methods involving this ratio was demonstrated by Fellegi and Sunter.

In the Fellegi-Sunter framework, a linkage rule assigns a probability of each classification decision (A_1, A_2 and A_3) to each outcome vector. The decision function corresponding to outcome vector $\underset{\sim}{x}$ is $d(\underset{\sim}{x}) = (P(A_1 \mid \underset{\sim}{x}), P(A_2 \mid \underset{\sim}{x}), P(A_3 \mid \underset{\sim}{x}))$. Acceptable rates of classification error for true non-matches and true matches are specified before linkage is conducted. We denote these pre-specified error rates by μ and λ respectively. Among the class of record linkage rules satisfying the relations $P(A_1 \mid U) \leq \mu$ and $P(A_3 \mid M) \leq \lambda$ for fixed values of μ and λ, Fellegi and Sunter define the optimal linkage rule as the rule that minimizes $P(A_2)$, the probability that a record pair will be classified as a possible link. The optimal rule has the form

$$
\begin{aligned}
d(\underline{x}^j) &= (1,0,0) \quad \text{if} \quad \omega^j > \tau_1 \\
d(\underline{x}^j) &= (P_\mu, 1 - P_\mu, 0) \quad \text{if} \quad \omega^j = \tau_1 \\
d(\underline{x}^j) &= (0,1,0) \quad \text{if} \quad \tau_2 < \omega^j < \tau_1 \qquad (2) \\
d(\underline{x}^j) &= (0, 1 - P_\lambda, P_\lambda) \quad \text{if} \quad \omega^j = \tau_2 \\
d(\underline{x}^j) &= (0,0,1) \quad \text{if} \quad \omega^j < \tau_2
\end{aligned}
$$

where $\tau_1 \geq \tau_2$, the "weight" ω^j is defined as $\omega^j = \log(R(\underline{x}^j))$ and P_μ and P_λ are positive constants in the interval $[0,1)$. (Refer to Fellegi and Sunter (1969) for full details.) Determination of τ_1 and τ_2 requires the estimation of classification error rates corresponding to various choices for these threshold values, underscoring the importance of accurate estimation of classification error rates in the Fellegi-Sunter framework.

Model-based estimates of classification error rates can be calculated using estimates of outcome probabilities for true matches and true non-matches. Let $\hat{P}(\underline{x} \mid M)$ and $\hat{P}(\underline{x} \mid U)$ denote estimates of the probabilities of outcome vector $\underline{x}$ for true matches and true non-matches and denote the ratio of these estimates by $\hat{R}(\underline{x})$. The model-based estimate of the classification error rate for true matches based on decision rule (2) is

$$\hat{\lambda} = \sum_{\underline{x} \in L(\tau_2)} \hat{P}(\underline{x} \mid M) + P_\lambda \sum_{\underline{x} \in Q(\tau_2)} \hat{P}(\underline{x} \mid M) \qquad (3)$$

where $L(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) < \tau_2\}$ and $Q(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_2\}$.

The model-based estimate of the classification error rate for true non-matches is

$$\hat{\mu} = \sum_{\underline{x} \in G(\tau_1)} \hat{P}(\underline{x} \mid U) + P_\mu \sum_{\underline{x} \in Q(\tau_1)} \hat{P}(\underline{x} \mid U) \qquad (4)$$

where $G(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) > \tau_1\}$ and $Q(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_1\}$.

2.2 A Model For Outcome Probabilities

Calculation of model-based classification error rate estimates requires estimation of $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ for each of the 2^K possible values of $\underline{x}$. The probability density function for $\underline{x}$ is a mixture of two probability densities given by

$$f(\underline{x}) = pP(\underline{x} \mid M) + (1 - p)\, P(\underline{x} \mid U), \qquad (5)$$

where p is the probability that a record pair chosen at random is a true match. The outcome probabilities depend on the frequency distributions of identifiers for entities represented on files A and B, as well as the probabilities that errors are introduced when identifiers are recorded on the files. Fellegi and Sunter (1969, pp. 1192-1194) describe a method of estimating agreement probabilities involving their definition in terms of frequency distributions and error probabilities. They recommend use of the method when prior information is available.

In the present paper, we consider situations in which the data on files A and B and the outcome vectors $\underline{x}^j$, $j = 1, 2, \ldots, N$, represent the only information available for estimation of outcome probabilities. A loglinear structure for the outcome probabilities is the most general parameterization. The saturated loglinear model for outcome probabilities for true matches is

$$
\begin{aligned}
\log(P(\underline{x} \mid M)) = {} & M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots \\
& + M(K)_{x_K} + M(1)\, M(2)_{x_1, x_2} + \ldots \\
& + M(K-1)\, M(K)_{x_{K-1}, x_K} + \ldots \\
& + M(1)\, M(2) \ldots M(K)_{x_1, x_2, \ldots, x_K}, \qquad (6)
\end{aligned}
$$

with the usual restrictions

$$\sum_{x_J} M(J)_{x_J} = 0, \quad J = 1, 2, \ldots, K,$$

$$\sum_{x_{J_1}} M(J_1) M(J_2)_{x_{J_1}, x_{J_2}} = \sum_{x_{J_2}} M(J_1) M(J_2)_{x_{J_1}, x_{J_2}} = 0,$$

$$\forall\ J_1, J_2, \qquad \textit{etc.},$$

as well as the restriction

$$\sum_{\underline{x}} P(\underline{x} \mid M) = 1.$$

The saturated model for $P(\underline{x} \mid U)$ is analogous.

If saturated loglinear models for $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ are employed, the density function includes $2^{K+1} - 1$ unknown parameters. It is not possible to identify all these parameters when no auxiliary information is available. In order to obtain a model that can be identified and to simplify the estimation problem, the assumption that the outcomes of comparisons for different data fields are independent is often employed. Under the assumption of independence, we denote the probabilities of agreement among record pairs that are true matches and true non-matches, respectively, by

$$m_k = P(x_k = 1 \mid M), \quad k = 1, 2, \ldots, K,$$

$$u_k = P(x_k = 1 \mid U), \quad k = 1, 2, \ldots, K.$$

Outcome probabilities can be written as

$$P(\underline{x} \mid M) = \prod_{k=1}^{K} m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\underline{x} \mid U) = \prod_{k=1}^{K} u_k^{x_k} (1 - u_k)^{1-x_k}.$$

This model involves $2 \cdot K + 1$ unknown parameters, namely $(\underline{m}, \underline{u}, p)$, where $\underline{m} = (m_1, m_2, \ldots, m_k)$, $\underline{u} = (u_1, u_2, \ldots, u_k)$. There are, of course, a number of intermediate models between the saturated model and the independence model. Methods that can be used to estimate the independence model are described in Section 3. Estimation of intermediate models is discussed in Section 4.

3. ESTIMATION UNDER INDEPENDENCE ASSUMPTION

3.1 Method of Moments

A methods of moments estimator of $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ can be employed in the case of independence. The estimator is based on a system of $2 \cdot K + 1$ equations that provide expressions for functionally independent moments of $\underline{x}$ in terms of the parameters. The equations are

$$E\left(\prod_{k \neq i}^{K} x_k\right) = pN \prod_{k \neq i}^{K} m_k + (1 - p) N \prod_{k \neq i}^{K} u_k,$$

$$i = 1, 2, \ldots, K$$

$$E(x_i) = pNm_i + (1 - p) Nu_i, \quad i = 1, 2, \ldots, K, \tag{7}$$

$$E\left(\prod_{k=1}^{K} x_k\right) = pN \prod_{k=1}^{K} m_k + (1 - p) N \prod_{k=1}^{K} u_k.$$

To obtain estimates of the parameters using the method of moments, it is necessary to solve the equations after expectations have been replaced by averages calculated using record pairs in C. The equation system for $K = 3$ was given by Fellegi and Sunter, who also derived a closed form solution that exists if some mild conditions are satisfied. Their paper included a word of caution concerning use of the method in the case of departures from independence. For $K > 3$, a closed form solution is not available but standard numerical methods can be used. Parameter estimates obtained using the method of moments are statistically consistent if the independence assumption is true.

3.2 Iterative Method

The iterative method was developed by record linkage practitioners. Although the method is not based on the probability distribution of the outcome vector, it does make use of the independence assumption. Application of the iterative method is described by several authors, including Newcombe (1988). Statistics Canada's record linkage software, CANLINK, is set up to facilitate use of the iterative method.

The method requires initial estimates of the agreement probabilities for true matches and non-matches. For true matches, guesses based on previous experience must be employed. To obtain initial estimates of agreement probabilities among record pairs that are true non-matches it is typically assumed that these probabilities are equal to the probabilities of agreement among record pairs chosen at random, namely that,

$$u_k = P(x_k = 1), \quad k = 1, 2, \ldots, K.$$

Suppose that $J(k)$ different values for data field k appear on file A and/or file B. Denote the frequencies of these values on file A by $f_{k1}, f_{k2}, \ldots, f_{kJ(k)}$ and denote the file B frequencies by $g_{k1}, g_{k2}, \ldots, g_{kJ(k)}$. For a particular value one, but not both, of the counts may be zero. The initial estimate of u_k is

$$\hat{u}_k^0 = \sum_{j=1}^{J(k)} (f_{kj} g_{kj})/N. \tag{8}$$

Given these probability estimates, initial sets of matches and non-matches, denoted by M^0 and U^0 respectively, are obtained using a decision rule

$$j \epsilon M^0 \quad \text{if} \quad \omega^j > \tau_1^0,$$

$$j \epsilon U^0 \quad \text{if} \quad \omega^j < \tau_2^0.$$

Next, frequency counts among record pairs in the sets M^0 and U^0 are used as new estimates of agreement probabilities. These estimates are used to obtain new sets of matches and non-matches and the iterative process is continued until consecutive estimates of agreement probabilities are sufficiently close.

In most applications, the assumption that the probability of agreement among record pairs that are true non-matches is equal to the probability of agreement among all record pairs is a good one and iteration does not lead to any important changes in estimates of non-match agreement probabilities. However, the first iteration often produces large changes in agreement probability estimates for true matches. Typically, there are no substantial changes at the second iteration.

It should be noted that the statistical properties of the iterative method are unclear. In practice, performance of the method will depend on the choice of the initial thresholds τ_1^0, τ_2^0. These thresholds are typically chosen subjectively. The simulations reported in Section 5 provide information about the effects of various initial thresholds.

4. RELAXING THE INDEPENDENCE ASSUMPTION - ESTIMATION USING ITERATIVE SCALING

Methods of estimation for latent variable models can be used to estimate agreement probabilities when the dependence between outcomes of comparisons for different matching fields is parameterized in terms of loglinear effects. Winkler (1989) and Thibaudeau (1989) have estimated agreement probabilities using loglinear models including all interaction terms up to third or fourth order to parameterize dependencies. The formulation presented here facilitates use of loglinear models including selected interactions. Match status can be considered a latent variable with two levels (true match and true non-match). Let $c_{0,\underline{x}}$ and $c_{1,\underline{x}}$ denote the numbers of true non-matches and true matches, respectively, with outcome vector $\underline{x}$ in a record linkage application involving K matching variables. These counts are, of course, unobservable since the value of the latent variable for each record pair is unknown. Instead, $c_{\underline{x}} = c_{0,\underline{x}} + c_{1,\underline{x}}$ is observed.

Using the parameterization of dependence in terms of loglinear effects and a saturated model for true matches, we can write

$$\begin{aligned}\log(c_{1,\underline{x}}/(Np)) = {} & M(0) + M(1)_{x_1} + M(2)_{x_2} + \dots \\ & + M(K)_{x_K} + M(1)M(2)_{x_1,x_2} + \dots \\ & + M(K-1)M(K)_{x_{K-1},x_K} + \dots \\ & + M(1)M(2)\dots M(K)_{x_1,x_2,\dots,x_K},\end{aligned}$$

with the usual restrictions. A similar expression for true non-matches is available. The latent variable model corresponding to these saturated loglinear models is

$$\begin{aligned}\log(c_{s,\underline{x}}/w_s) = {} & G(0) + Z_s + G(1)_{x_1} + \dots \\ & + G(K)_{x_K} + ZG(1)_{s,x_1} + \dots + ZG(K)_{s,x_K} \\ & + \dots + G(1)G(2)\dots G(K)_{x_1,x_2,\dots,x_K} \\ & + ZG(1)G(2)\dots G(K)_{s,x_1,x_2,\dots x_K},\end{aligned}$$

where the index s has value zero for true non-matches and one for true matches, $w_0 = (1-p)N$ and $w_1 = pN$. The parameters are analogous to the parameters of a saturated loglinear model for a contingency table of dimension 2^{K+1}. The usual restrictions apply. For example, the term $ZG(1)_{s,x_1}$ represents the interaction of the latent variable and the first matching variable and

$$\sum_s ZG(1)_{s,x_1} = \sum_{x_1} ZG(1)_{s,x_1} = 0.$$

This model conforms to the general latent variable model of Haberman (1979, p. 561). Additional restrictions must be imposed to identify and estimate the parameters. For simplicity, we will consider only hierarchical models. In addition, we restrict attention to models that allow all non-zero effects to interact with the latent variable.

In subsequent discussion we will denote latent variable models using symbols $G(1), G(2), \dots$, loglinear models for true matches using $M(1), M(2), \dots$ and loglinear models for true non-matches using $U(1), U(2), \dots$. In the case of four matching variables, for example, the model $G(1)G(2), G(3), G(4)$ is a latent variable model including a general level term, main effects for all four matching variables and a term for the interaction of matching variables one and two, as well as a main effects term for the latent variable (the interaction of the general level term and the latent variable), terms for the interaction of each matching variable and the latent variable and a term for the interaction of matching variables one and two and the latent variable. The model includes 12 parameters that must be estimated. The number of parameters that must be estimated in one of the latent variable models considered here is twice the number of parameters in the corresponding loglinear model.

The iterative scaling method of Haberman (1976) can be used to estimate latent variable models. The Haberman estimation method operates by raking tables that contain estimated counts for each outcome among true matches and true non-matches. Denote the estimated counts for outcome vector $\underline{x}$ after i iterations of the Haberman algorithm by $\hat{C}^i_{1,\underline{x}}$ and $\hat{C}^i_{0,\underline{x}}$ for true matches and true non-matches, respectively. Starting values $\hat{C}^0_{1,\underline{x}}$ and $\hat{C}^0_{0,\underline{x}}$ can be constructed using estimates of agreement probabilities and the proportion of true matches obtained under the independence assumption. Each iteration of the algorithm involves a series of raking operations on the current table for true matches and the analogous rakes on the current table for true non-matches. Using the notation for hierarchical models introduced above, a set a raking operations is performed for each of the interaction terms that define the model. For four matching variables and the model $G(1)G(2), G(3)G(4)$, two sets of raking operations are performed - one for the $G(1)G(2)$ interaction and a second for the $G(3)G(4)$ interaction. For each iteraction, one raking operation is performed for every level of the corresponding classification variable. Let S_{gl} denote the set of outcome vectors at level l of term g. The raking operation on the table of true matches at iteration i for level l of term g involves computation of

$$\gamma_{1,\underline{x}} = c_{\underline{x}}\hat{c}^{i-1}_{1,\underline{x}} / (\hat{c}^{i-1}_{1,\underline{x}} + \hat{c}^{i-1}_{0,\underline{x}}),$$

$$\hat{c}^i_{1,\underline{x}} = \hat{c}^{i-1}_{1,\underline{x}} \sum_{\underline{x}\epsilon s_{gl}} \gamma_{1,\underline{x}} \Big/ \sum_{\underline{x}\epsilon s_{gl}} \hat{c}^{i-1}_{1,\underline{x}}, \quad \forall\, \underline{x}\epsilon S_{gl}.$$

The algorithm is terminated when changes between estimated counts for consecutive iterations are smaller than a given tolerance.

Haberman (1976) notes that the iterative scaling algorithm may converge to a local maximum of the likelihood function rather than to the maximum likelihood estimate. Experiments with different starting values using data sets employed in the evaluation reported in Section 5 did not yield any examples of this problem.

5. COMPARISON OF ESTIMATION METHODS - SYNTHETIC DATA

In this section, the results of comparisons of the estimation methods described in Section 3 and Section 4 are presented. The comparisons involved application of each approach to a series of synthetic data sets generated using Monte Carlo methods.

Synthetic data records containing four personal identifiers (family name, middle initial, given name, date of birth) were employed. Information on possible values of each identifier, as well as their relative frequencies, was taken from the Canadian Mortality Data Base for 1988. This database, which is frequently used in health applications of record linkage, contains a separate record for each individual death.

The independence assumption was violated among true matches in each synthetic data set. Information on the frequency of outcome vectors for true matches obtained from various record linkage projects conducted by the Canadian Center for Health Information at Statistics Canada was used during data generation. Most of the projects involved matching a cohort file to the Canadian Mortality Data Base. The frequency of each outcome vector among the true matches is shown in Table 1. The dependence in these data is clear. Although approximately 88.3% of the true matches agree on given name, the probability of agreement on given name given disagreement on middle initial and agreement on family name and birth year is only 381/1366 - about 27.9%. The value of the likelihood ratio test statistic for the independence hypothesis is 3604. This value is very extreme relative to the chi-square reference distribution with 10 degrees of freedom. (Note that one degree of freedom is lost due to the zero count for the cell (1,0,0,0).)

For each synthetic data set, file A records were generated by selecting identifiers according to relative frequencies in the 1988 Canadian Mortality Data Base. In order to simplify the data generation process, the choice of family names was restricted to the 100 most common non-francophone family names and the 100 most common francophone family names found on the 1988 file. The choice of given name was restricted to the 50 most common francophone given names and the 50 most common non-francophone given names. All name choices excluded typographical variations. All middle initials and birth years found on the 1988 file were considered. Records with anglophone given names were more likely to receive an anglophone family name than records with francophone given names (reflecting the distribution of names in the Canadian population). Otherwise, identifiers were selected independently.

Table 1

Outcome Frequencies, Set of True Matches, Synthetic Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Frequency	
Given Name	Middle Initial	Family Name	Birth Year	Count	Percentage
0	0	0	0	7	0.03
0	0	0	1	33	0.12
0	0	1	0	125	0.45
0	0	1	1	985	3.54
0	1	0	0	5	0.02
0	1	0	1	39	0.14
0	1	1	0	202	0.73
0	1	1	1	1,848	6.65
1	0	0	0	0	0.0
1	0	0	1	13	0.05
1	0	1	0	50	0.18
1	0	1	1	381	1.37
1	1	0	0	44	0.16
1	1	0	1	451	1.62
1	1	1	0	1,751	6.30
1	1	1	1	21,860	78.65
			Total	27,794	100

The starting point for file B was an exact copy of file A. Each file B record was a true match with exactly one file A record. To introduce dependence among true matches, an outcome vector was drawn from the frequency distribution in Table 1 for each file B record. Identifiers corresponding to zeroes in the outcome vector were re-selected. Consequently, the set of outcome vectors for true matches was a sample from the Table 1 distribution. The synthetic data sets also included mild departures from the independence assumption for true non-matches since the selection of given and family names was not completely independent.

Each set of simulation results reported subsequently is based on 50 Monte Carlo trials. Each trial involved generation of files A and B of size 500, estimation of $\underline{m}$ and $\underline{u}$, determination of thresholds corresponding to various model-based classification error rate estimates and calculation of actual error rates corresponding to the thresholds. The same series of 50 synthetic data sets was used for each set of simulations. Note that the set C contains 250,000 record pairs including 249,500 true non-matches for each Monte Carlo trial. In order to reduce computing time required by the simulations, only 49,500 true non-matches were used for each trial. (A small scale test was conducted to verify that reducing the number of true non-matches had a negligible effect on the estimated agreement probabilities.) True non-matches were removed from C by dividing files A and B into five corresponding blocks of size 100 and excluding record pairs involving records from blocks that did not correspond.

The method of moments equation system was solved using a variation of Newton's method that is described in detail in Moré *et al.* (1980). Computer code from IMSL (1987) was employed. Agreement probabilities of 0.9 for true matches and 0.1 for true non-matches for all matching fields were used as starting values for the solution of the equation system. The method did not appear sensitive to starting values.

The properties of the iterative method depend on the definitions of the initial sets of matches and non-matches, M^0 and U^0. Recall that, given initial probabilities, record pairs are classified according to

$$j \epsilon M^0 \quad \text{if} \quad \omega^j > \tau_1^0,$$

$$j \epsilon U^0 \quad \text{if} \quad \omega^j < \tau_2^0.$$

When the iterative method was implemented for the simulations reported here, τ_2^0 was set equal to τ_1^0. For each Monte Carlo trial, τ_1^0 was determined such that

$$\hat{P}(j \epsilon U \mid \omega^j > \tau_1^0) + \gamma \cdot \hat{P}(j \epsilon U \mid \omega^j = \tau_1^0) = \mu^0,$$

for some $\gamma \epsilon [0,1)$, where the estimated probabilities are based on the initial iterative estimates of $\underline{u}$. Record pairs with weight τ_1^0 were classified in M^0 with probability γ. That is, the initial set of matches used by the iterative method was chosen to correspond to an estimated classification error rate of μ^0 for true non-matches. Starting values for m_k, $k = 1, 2, \ldots, 4$, were set to 0.9.

The zero count in Table 1 (agreement on given name, disagreement on all other identifiers) was treated as a structural zero during data generation. Among loglinear models involving no more than six parameters the model that gives the best fit to the Table 1 data is $M(1)M(2)$, $M(3)$, $M(4)$. This model, involving dependence for outcomes of comparisons for given name and middle initial, does not fit particularly well. The likelihood ratio test statistic for lack of fit is 57.95 - an extreme value relative to the chi-square reference distribution with 9 degrees of freedom. The latent variable model $G(1)G(2)$, $G(3)$, $G(4)$ was estimated for each synthetic data set using iterative scaling. This model fit the synthetic data sets somewhat better than the model $M(1)M(2)$, $M(3)$, $M(4)$ fit the true match data. The largest lack of fit test statistic among the fifty synthetic data sets was 25.03 and the model was rejected only ten times at the 5% level of significance.

Averages of classification error rate estimates obtained using the synthetic data sets and the corresponding Monte Carlo standard errors are reported in Table 2 for true non-matches and Table 3 for true matches. After multiplication by 99, the error rates for true non-matches represent numbers of misclassified true non-matches divided by numbers of true matches. Results are given for the method of moments and iterative scaling, as well as the iterative method with μ^0 = 0.0000625, 0.00025 and 0.001. The biases in estimated error rates for true non-matches are generally small. The iterative method with μ^0 = 0.001

Table 2

Classification Error Rates, True Non-matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate (× 99)	Actual Rate (× 99)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0188 (0.0008)	0.0208 (0.0008)	0.0208 (0.001)	0.0207 (0.001)	0.0195 (0.001)
0.04	0.0381 (0.001)	0.0408 (0.0013)	0.0407 (0.0016)	0.0405 (0.0016)	0.0397 (0.0016)
0.06	0.057 (0.0012)	0.0626 (0.0015)	0.0615 (0.0018)	0.0602 (0.0019)	0.059 (0.0018)
0.08	0.076 (0.0015)	0.0855 (0.0017)	0.0838 (0.0019)	0.0804 (0.0022)	0.0785 (0.0019)
0.10	0.095 (0.0019)	0.1086 (0.0021)	0.1061 (0.0022)	0.1007 (0.0026)	0.0978 (0.0021)

Table 3

Classification Error Rates, True Matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0580 (0.0013)	0.1179 (0.0041)	0.0507 (0.0014)	0.0149 (0.0008)	0.025 (0.0012)
0.04	0.0773 (0.0014)	0.1362 (0.004)	0.0735 (0.0012)	0.0359 (0.0018)	0.0455 (0.0016)
0.06	0.0966 (0.0014)	0.1542 (0.0038)	0.0954 (0.0012)	0.0660 (0.0014)	0.0646 (0.0018)
0.08	0.1159 (0.0014)	0.1722 (0.0036)	0.1165 (0.0012)	0.0866 (0.0017)	0.0841 (0.0019)
0.10	0.1348 (0.0014)	0.1904 (0.0035)	0.1319 (0.0014)	0.1025 (0.002)	0.1043 (0.002)

provides the best estimates, followed by iterative scaling. For true matches the performance of the iterative method is very sensitive to the choice of μ^0. Although the iterative method performs well for $\mu^0 = 0.001$, the biases for $\mu^0 = 0.0000625$ and $\mu^0 = 0.00025$ are substantial. Estimates of classification error rates for true matches obtained using the method of moments also include large biases. Biases in estimates based on iterative scaling are relatively small.

Table 4

Classification Error Rates, True Non-matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate (× 99)	Actual Rate (× 99)	
	Method of Moments	Iter. Scaling
0.02	0.0189 (0.0008)	0.0194 (0.001)
0.04	0.0385 (0.0011)	0.0396 (0.0016)
0.06	0.0577 (0.0013)	0.0589 (0.0019)
0.08	0.0767 (0.0016)	0.0785 (0.002)
0.10	0.0957 (0.002)	0.0978 (0.0021)

The information in Tables 4 and 5 is based on a series of synthetic data sets generated using a modified version of Table 1. Expected values of Table 1 cell counts under the model $M(1)M(2)$, $M(3)$, $M(4)$ were used for data generation. The biases in model-based classification error rate estimates obtained using the method of moments are greatly reduced using the latent variable model $G(1)G(2)$, $G(3)$, $G(4)$ estimated using iterative scaling, particularly for true matches.

Table 5

Classification Error Rates, True Matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate	
	Method of Moments	Iter. Scaling
0.02	0.0553 (0.0014)	0.0208 (0.0011)
0.04	0.0747 (0.0014)	0.0415 (0.0016)
0.06	0.094 (0.0014)	0.0608 (0.0018)
0.08	0.1134 (0.0014)	0.0805 (0.002)
0.10	0.1325 (0.0015)	0.1007 (0.002)

6. COMPARISON OF ESTIMATION METHODS - REAL DATA

Results of comparisons of the three estimation methods using data from a record linkage application are presented in this section. Two data files used in empirical work reported by Fair and Lalonde (1987) were employed. The first file contained information on Ontario miners obtained from the Workmen's Compensation Board. The second file included information from the Canadian

Mortality Data Base (CMDB) for individual deaths during the period 1964 to 1977 inclusive. The miners' file included only those records with a valid social insurance number. The second file contained records that had survived an initial comparison exercise designed to eliminate records with no similarity to any of the records on the miners' file. The vital status of each miner at the end of 1977 had been classified as "confirmed dead", "confirmed alive" or "lost to follow-up" based on a previous linkage, combined with thorough follow-up procedures, including manual review. Records on the miners' file for individuals "confirmed dead" included the CMDB death registration number. More information on the construction of the files and the procedures used to determine true link status can be found in Fair and Lalonde.

Four identifiers - given name, NYSIIS code of mother's maiden name, day of birth and birth month - were chosen as matching fields for the comparison. Records on the miners' file with vital status "lost to follow-up" were eliminated. After records with missing values for at least one matching field or for birth year were also removed, file A (based on the miners' file) contained 45,638 records and file B (based on the CMDB) included 24,597 records. Restricting comparisons of the two files to pairs of records with the same NYSIIS representation of family name and birth years differing by at most one, there were 26,500 true non-matches and 2063 true matches.

Frequencies of outcomes among true matches and true non-matches are shown in Table 6. All loglinear models corresponding to a non-saturated latent variable model (that is, all models with fewer than eight parameters) are rejected by the frequency data for true non-matches at a very low level of significance. Among models with fewer than eight parameters the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ corresponds to the lowest likelihood ratio test statistic for lack of fit - 35.29. The model $M(1)$, $M(2)M(4)$, $M(3)M(4)$ provides an adequate fit to the true match data (likelihood ratio test statistic of 10.29).

Agreement probability estimates were computed using the method of moments, the iterative method and iterative scaling using the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$. The likelihood ratio test statistic for the independence model corresponding to the method of moments estimator is 108 (six degrees of freedom). The independence model is rejected by the data at a very low significance level. In contrast, the likelihood ratio test statistic for the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$ is 1.44 (two degrees of freedom), suggesting an adequate fit. Model-based estimates of classification error rates corresponding to each set of probability estimates were calculated for various thresholds. Actual classification error rates are compared to model-based estimates for true non-matches in Table 7 and true matches in Table 8. The error rates for true non-matches have been rescaled so that the number of true matches is in the denominator.

Table 6

Outcome Frequencies, Real Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Count	
Given Name	NYSIIS of Mother's Maiden Name	Day of Birth	Birth Month	True Matches	True Non-Matches
0	0	0	0	4	22,100
0	0	0	1	3	888
0	0	1	0	11	2,322
0	0	1	1	128	211
0	1	0	0	3	199
0	1	0	1	7	19
0	1	1	0	27	27
0	1	1	1	242	13
1	0	0	0	9	576
1	0	0	1	10	32
1	0	1	0	52	94
1	0	1	1	392	4
1	1	0	0	27	13
1	1	0	1	32	1
1	1	1	0	115	0
1	1	1	1	1,001	1
			Total	2,063	26,500

Model-based classification error rate estimates obtained using the iterative method are very inaccurate, particularly for true non-matches, regardless of the value of μ^0. Error rate estimates obtained using iterative scaling are slightly less accurate than estimates based on the method of moments for true matches. However, they are considerably more accurate than method of moments estimates for true non-matches.

Some words of caution are necessary. Even though the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ does not adequately describe the dependencies among true non-matches, the iterative scaling algorithm obtained a good fit using an estimate of the proportion of matched records (0.0747) that differs somewhat from the true value (0.0722). A similar fit can also be obtained using the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ and an estimate of 0.077 for the proportion of matches. Error rate estimates based on the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ are no better than estimates obtained using the method of moments.

Table 7

Classification Error Rates, True Non-matches, Real Data

Estimated Rate (× 12.84)	Actual Rate (× 12.84)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0368	1.311	0.1859	0.186	0.0339
0.04	0.0796	1.314	0.1888	0.193	0.0649
0.06	0.1224	1.317	0.1917	0.1967	0.0684
0.08	0.1573	1.323	0.1990	0.1994	0.1106
0.10	0.1863	1.333	0.60	0.4066	0.1282

Table 8

Classification Error Rates, True Matches, Real Data

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0166	0.0141	0.0193	0.0225	0.0105
0.04	0.0318	0.0264	0.029	0.0278	0.0263
0.06	0.0598	0.0383	0.0472	0.0326	0.0529
0.08	0.0782	0.0416	0.1372	0.0488	0.0784
0.10	0.0966	0.045	0.1393	0.1371	0.0958

7. CONCLUSIONS

In this paper, the issue of classification error rate estimation for record linkage has been discussed. The Fellegi-Sunter framework provides for the calculation of classification error rate estimates using estimates of agreement probabilities. These model-based estimates typically have poor properties in practice. It has been demonstrated that their properties can be improved through careful estimation of agreement probabilities. Three estimation methods have been evaluated using synthetic data as well as information from a real application.

For two of the three methods, the assumption that outcomes of comparisons for different data fields are independent was employed. This assumption was not valid for either the synthetic data or the real data. The synthetic data included strong dependencies for true matches and minor dependencies for true non-matches. Dependencies in the real data were particularly strong for true non-matches. Classification error rate estimates obtained using the method of moments, which relies on the assumption of independence, included substantial bias for synthetic data and were relatively inaccurate for real data. The magnitude of the bias in classification error rate estimates for synthetic data obtained using the iterative method depended on the definition of an initial set of matches. Although some definitions of the initial set of matches led to relatively small biases, others produced estimates with biases much larger than those obtained using the alternative methods. For the real data, all the definitions of the initial set of matches considered led to very inaccurate error rate estimates. There are no mathematical rules available for the choice of an initial set of matches for the iterative method. The results in this paper provide no evidence to recommend its use.

The third method relies on a parameterization of dependencies between outcomes of comparisons for different data fields using loglinear effects. Under this parameterization, estimates of agreement probabilities that do not rely on the independence assumption can be obtained through use of the iterative scaling method to estimate the parameters of a latent variable model. For the synthetic data sets with lack of independence, model-based classification error rate estimates obtained using iterative scaling included much smaller biases than estimates based on the independence assumption. Although the latent variable model fit most synthetic data sets better than a model based on the independence assumption, it sometimes exhibited significant lack of fit. When the synthetic data was modified to improve the fit of the latent variable model, there was no evidence of bias in model-based classification error rate estimates. The real data included important departures from independence for both true matches and true non-matches. Model-based error rate estimates obtained using iterative scaling were slightly less accurate than estimates based on the method of moments for true matches and considerably more accurate for true non-matches.

The results reported here indicate that properties of model-based classification error rates estimates can be improved using an appropriate estimator of agreement probabilities. Latent variable models and iterative scaling provide a method of incorporating dependencies between outcomes of comparisons for different data fields during estimation of agreement probabilities.

ACKNOWLEDGEMENTS

The authors would like to thank William Winkler for providing the computer code that was the basis of the iterative scaling estimation program used to obtain our results, as well as Fritz Scheuren and three anonymous referees for comments on a earlier version of this paper that led to a significant improvement in both the content and the presentation. Thanks are also due to Martha Fair and Pierre Lalonde for making available the Ontario miners' data and the outcome frequency data for true matches.

REFERENCES

BARTLETT, S., KREWSKI, D., WANG, Y., and ZIELINSKI, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.

BELIN, T.R. (1990). A proposed improvement in computer matching techniques. In *Statistics of Income and Related Administrative Record Research*: 1988-1989, U.S. Internal Revenue Service, 167-172.

BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.

FAIR, M.E., and LALONDE, P. (1987). Missing identifiers and the accuracy of individual follow-up. *Proceedings: Symposium on Statistical Uses of Administrative Data, Statistics Canada*, 95-107.

FAIR, M.E., NEWCOMBE, H.B., and LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Research report, Atomic Energy Control Board.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.

IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.

MORÉ, J., GARBOW, B., and HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 149–163
Statistics Canada

Robust Joint Modelling of Labour Force Series of Small Areas

D. PFEFFERMANN and S.R. BLEUER[1]

ABSTRACT

In this article we report the results of fitting a state-space model to Canadian unemployment rates. The model assumes an additive decomposition of the population values into a trend, seasonal and irregular component and separate autoregressive relationships for the six survey error series corresponding to the six monthly panel estimators. The model includes rotation group effects and permits the design variances of the survey errors to change over time. The model is fitted at the small area level but it accounts for correlations between the component series of different areas. The robustness of estimators obtained under the model is achieved by imposing the constraint that the monthly aggregate model based estimators in a group of small areas for which the total sample size is sufficiently large coincide with the corresponding direct survey estimators. The performance of the model when fitted to the Atlantic provinces is assessed by a variety of diagnostic statistics and residual plots and by comparisons with estimators in current use.

KEY WORDS: Design variance; Kalman filter; Panel survey; Rotation bias; State-space model.

1. INTRODUCTION

A time series model for survey data is the combination of two distinct models. The "census model" describing the evolution of the finite population values over time and the survey errors model representing the time series relationships between the survey errors of the survey estimators. There are at least four main reasons for wishing to model the raw survey estimators:

(a) The model based estimators of the population values resulting from the modelling process have in general smaller variances than the survey estimators, particularly in small areas where the sample sizes are small.

(b) The model we employ yields estimators for the seasonal effects and for the variances of these estimators as a by-product of the estimation process.

(c) The model can be used to forecast the population values, the trend and the seasonal components for time periods beyond the sample time period for which the direct survey estimators are available. Such forecasts are important when assessing the performance of the model and for policy decision making.

(d) The model can be used to detect turning points in the level of the series and assess their significance. (Work on this problem will be addressed in a separate article).

The methodology described in this article integrates the methodologies presented in Pfeffermann and Burck (1990) and Pfeffermann (1991) with some new modifications and extensions. The main features of the model are as follows:

1. The model decomposes the population values into the unobservable components of trend, seasonality and irregular terms. Smoothed predictors of the three components (and hence of the population values) based on all the available data, and standard errors of the prediction errors are obtained straightforwardly by application of the Kalman filter. The standard errors are modified to account for the extra variation induced by the use of estimated parameter values.
2. The model uses the distinct monthly panel estimators as input data. The use of the panel estimators has two important advantages over the use of the mean estimators: (i) It identifies better the time series model holding for the survey errors by analysing contrasts between the panel estimators, (ii) It yields more efficient estimators for the model parameters and hence better predictors for the unobservable model components.
3. The model accounts for changes in the variances of the survey errors over time and for possible rotation group effects.
4. The model can be applied simultaneously to the panel estimators in separate small areas. The census model is extended in this case to account for the cross-correlations between the unobservable components of the population values operating in these areas.
5. A modification to ensure the robustness of the small area estimators against possible model breakdowns is incorporated into the model equations. The modification consists of constraining the model based estimators of aggregates of the population values over a group of small areas for which the total sample size is sufficiently large to coincide with the corresponding aggregate survey estimators. As a result, sudden changes in the level of the series are reflected in the model based estimators with no time lag.

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905; S.R. Bleuer, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The model and the robustness modifications are described in more detail in section 2. Empirical results obtained when fitting the model to the four Atlantic provinces of Canada are presented in section 3. Section 4 contains a short summary with suggestions for extension of the analysis.

Before concluding this section we mention that in the U.S., the state unemployment estimates are produced for most of the states based on time series models which have a similar structure to the model used in our study. See Tiller (1992) for details. A major difference between the two models is that in the U.S., the model postulated for the population values includes also explanatory variables so that the trend and the seasonal component only account for the trend and seasonal variations not accounted for the explanatory variables. The models fitted to the survey errors are like in our case of the ARIMA type and they likewise account for changes in the variances of the survey errors. They are otherwise different because of the very different sample rotation schemes used in the two countries. Another notable difference between the two models is that in the U.S., the models are fitted to each state separately and the input data consist of only the mean survey estimates, that is, one observation for every month. As a result, the models do not account for rotation group biases.

2. A STATE-SPACE MODEL FOR CANADA UNEMPLOYMENT SERIES

2.1 The Canadian Labour Force Survey

Data on unemployment are collected as part of the Labour Force Survey (LFS) carried out by Statistics Canada. The Canadian LFS is a rotating monthly panel survey by which every new sampled panel of households is retained in the sample for six successive months before being replaced by another panel from the same PSU's or strata. The PSU's are defined by geographic locations (city blocks or urban centers in the urban regions and groups of enumeration areas in the rural regions). The strata are homogeneous groups of PSU's defined by geographic locations such as city tracts, census subdivisions and enumeration areas. In the urban regions, (about 2/3 of the sample), every PSU is represented in only one panel. In the rural regions, the PSU's are represented in all the panels but with different enumeration areas in different panels. As a result, the separate panel estimators can be assumed to be independent, a property validated and utilized in other studies, see *e.g.* Lee (1990). For a recent report describing the design of the LFS and the construction of the direct survey estimators, the reader is referred to Singh *et al.* (1990).

2.2 The Census Model

In what follows we consider a single small area. In section 2.4 we consider joint modelling of the panel estimates in a group of small areas. The model postulated for the population values is the Basic Structural Model (BSM) which consists of the following set of equations.

$$Y_t = L_t + S_t + \epsilon_t; \quad L_t = L_{t-1} + R_{t-1} + \eta_{Lt};$$

$$R_t = R_{t-1} + \eta_{Rt}; \quad \sum_{j=0}^{11} S_{t+j} = \eta_{St}. \tag{2.1}$$

In (2.1) Y_t is the population value ("true" unemployment rate) at time t, L_t, is the trend level, R_t is the increment, S_t the seasonal effect and ϵ_t the irregular term assumed to be white noise with zero mean and variance σ_ϵ^2. Thus, the first equation in (2.1) postulates the classical decomposition of a time series into a trend, seasonal and irregular components. This decomposition is inherent in the commonly used procedures for seasonal adjustment, see *e.g.* Dagum (1980). Notice however that in the present case the series $\{Y_t\}$ is itself unobservable. The series $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ are independent white noise disturbances with mean zero and variances σ_L^2, σ_R^2 and $\sigma_S^2 \times g(t)$ respectively. Hence, the second and third equations of (2.1) define a local approximation to a linear trend whereas the last equation models the evolution of the seasonal effects such that the sum of every 12 successive effects fluctuates around zero. Notice that the variances of the error terms η_{St} are time dependent. The functions $g(t)$ are specified at the end of section 3.1.

The theoretical properties of the BSM in comparison to other models are discussed in Harrison and Stevens (1976), Harvey (1984) and Maravall (1985). Empirical results illustrating the performance of the model are shown in Harvey and Todd (1983), Morris and Pfeffermann (1984) and Pfeffermann (1991). Although more restricted than the family of ARIMA models, the BSM is now recognized as being flexible enough to approximate the behaviour of many diverse time series.

2.3 The Survey Errors Model

The model holding for the survey errors was identified initially by analyzing separately the pseudo error series $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $t = 1, \ldots, N$, where $y_t^{(j)}$ is the estimator of Y_t based on j-th panel $j = 1, \ldots, 6$, (the panel surveyed for the j-th successive month) and $\bar{y}_t = \sum_{j=1}^{6} y_t^{(j)}/6$ is the mean estimator. Notice that $(y_t^{(j)} - \bar{y}_t) = (e_t^{(j)} - \sum_{j=1}^{6} e_t^{(j)}/6)$, where $e_t^{(j)} = (y_t^{(j)} - Y_t)$ are the true survey errors. Thus, the notable feature of the contrasts $(y_t^{(j)} - \bar{y}_t)$ is that they are functions of only the survey errors irrespective of the model holding for the population values.

There are two prior considerations in the choice of a model for the survey errors:

(a) The model should account for possible rotation group biases or more generally, allow for different means for the survey errors of different panels.

(b) The model should account for changes in the variances of the survey errors over time.

Rotation group biases may arise from providing different information on different rounds of interview, depending on the length of time that respondents are included in the sample, or on the method of data collection, say, whether by telephone or by home interview. (In the Canadian LFS, the first panel is interviewed by home visits, the other panels are interviewed by telephone). Another possible reason for differences between the panel survey error means is differences in the nonresponse patterns across the panels. See Pfeffermann (1991) for further discussion with references to earlier studies on this problem.

Changes in the variances of the survey errors over time occur when the variances are function of the level of the series. Indeed, as revealed by figure 1 in section 3, the estimates of the standard deviations of the survey errors are subject to seasonal effects with a seasonal pattern that follows the seasonal pattern of the population values. Another possible explanation for changes in the variances of the survey errors is changes in the sampling design. For example, the overall sample size of the Canadian LFS was reduced in 1985-1986 from 55,000 households to 48,000 households. This reduction in the sample size was associated with other changes in the design. See Singh *et al.* (1990) for details.

Application of simple model estimation and diagnostic procedures to the pseudo survey errors suggest a 3rd order autoregressive (AR) model for the standardized survey errors $\tilde{e}_t^{(j)} = (e_t^{(j)} - \beta_j)/SD(e_t^{(j)})$, *i.e.*

$$\tilde{e}_t^{(j)} = \phi_{j1}\,\tilde{e}_{t-1}^{(j-1)} + \phi_{j2}\,\tilde{e}_{t-2}^{(j-2)} + \phi_{j3}\,\tilde{e}_{t-3}^{(j-3)} + u_t^{(j)},\ j = 1, \ldots, 6, \quad (2.2)$$

where $\beta_j = \mathrm{E}(e_t^{(j)})$ are the rotation group biases, $SD(e_t^{(j)})$ are the design standard deviations and $u_t^{(j)}$ are independent white noise with mean zero and variances σ_j^2. It is assumed that $\sum_{j=1}^{6} \beta_j = 0$ which implies that the mean survey estimator, $\bar{y}_t$, is unbiased. See Pfeffermann (1991) for discussion on the need to constraint the bias coefficients. Subsequent analysis when fitting the combined model defined by (2.1) and (2.2) (see section 2.4) validates this model with the further observation that the coefficients $(\phi_{j1}, \phi_{j2}, \phi_{j3})$ can be assumed to be equal for $j = 4, 5, 6$. Furthermore, for the first panel an AR(1) model already gives a good fit whereas for the second and third panel an AR(2) model is appropriate although with different coefficients. These relationships hold for each of the four Atlantic provinces.

One of the referees of this article raised the question of whether the AR(3) model defined by (2.2) is flexible enough to account for the panel estimates correlations at high lags which are believed to be high because of "PSU effects". As mentioned in section 2.1, panels rotating out of the sample are replaced by panels from the same PSU's and it usually takes several years before a PSU is exhausted and replaced by a neighbouring PSU. Lee (1990) presents two sets of panel estimates correlations for the Canadian LFS. The first set, denoted by ρ_j, are the correlations between estimates produced from the same panel so that j ranges from 1 to 5. The second set, denoted by γ_j, are the correlations between estimates produced from a panel and its predecessor so that j ranges from 1 to 11. The ρ-correlations are generally high as expected but it should be emphasized that they are lower for the unemployment series than for the employment series, demonstrating the high mobility of the unemployment Labour Force. The γ-correlations are much smaller than the ρ correlations but as mentioned by the author, the computation of these correlations is much less reliable and their behavior is somewhat fuzzy showing occasionally an increasing trend. We computed the serial correlations based on the models (2.2) with the ϕ-coefficients replaced by their estimated values and found in general a close fit to the ρ-correlations at all the lags from 1 to 5. The correlations at higher lags are different from the corresponding γ-correlations but interesting enough, they are in most cases higher and always decrease as j increases.

Another question related to the model (2.2) raised by the referees is whether one could apply the log transformation to the raw data for stabilizing the survey error variances, rather than modelling the standardized errors. There are two main reasons for not using the log transformation in our case. Foremost, the use of this transformation would imply a multiplicative decomposition for the population unemployment rates which is counter to common practice of postulating an additive decomposition. In Statistics Canada the unemployment rates in the two larger provinces out of the four considered in our study are deseasonalized by postulating the additive decomposition. In the U.S. the models fitted to the state unemployment series likewise postulate an additive decomposition. See Tiller (1992). The second reason is that changes in the survey error variances may result from charges in the sampling design and in particular, from changes in the sample sizes. Such changes cause discrete shifts in the variances which cannot be handled effectively by the log transformation. As noted also by one of the referees, transforming the data has the drawback of producing nonlinearity in aggregating the estimates over the panels and/or the small areas.

The model defined by (2.2) satisfies the two prior considerations discussed above. The actual application of the model requires however two modifications:

1. For the first three panels there is not a long enough history to permit the fitting of an AR(3) model. For example, the survey error $e_t^{(1)}$ corresponds to the panel which is in the sample for the first time. In order to overcome this problem, we replace the missing survey errors by the survey errors corresponding to the panels previously selected from the same PSU's or strata. For example, the AR(2) model fitted to $\bar{e}_t^{(2)}$ is

$$\tilde{e}_t^{(2)} = \phi_{21}\,\tilde{e}_{t-1}^{(1)} + \phi_{22}\,\bar{e}_{t-2}^{(6)} + u_t^{(2)}. \qquad (2.3)$$

   Notice that the panel surveyed for the second time at month t replaces at time $(t-1)$ the panel surveyed for the sixth time at month $(t-2)$ so that both panels represent the same PSU's or strata. The use of surrogate survey errors in the case of the first three panels may explain the different models identified for these panels as compared to the model identified for the other three panels.

2. The true standard deviations of the survey errors are unknown whereas the survey estimates of the standard deviations are themselves subject to sampling errors. To overcome this problem, we use smoothed values of the estimated standard deviations, obtained by fitting the relationship

$$(\widetilde{SD})_t = \hat{\gamma}(\widehat{SD})_{t-1} + \hat{\gamma}_0\,t + \sum_{i=1}^{12} \hat{\gamma}_i D_{it}, \qquad (2.4)$$

   with the γ-coefficients estimated by ordinary least squares. The notation $(\widehat{SD})_t$ defines the raw, unsmoothed estimate of the design standard deviation of the mean survey estimator, $\bar{y}_t$, at month t and $\{D_{it}\}$ are dummy variables accounting for monthly seasonal effects so that $D_{it} = 1$ when $t = 12k + i$, $k = 0, 1, \ldots, i = 1, \ldots, 12$ and $D_{it} = 0$ otherwise. The smoothed standard deviations of the panel survey errors are obtained as $\widetilde{SD}(e_t^{(j)}) = \sqrt{6}(\widetilde{SD})_t$. The latter estimates are used as surrogates for the true, unknown, standard deviations.

2.4 State-space Representation and Estimation of the Model Holding for the Survey Estimators

It follows from (2.1) that the panel estimators can be modeled as

$$y_t^{(j)} = L_t + S_t + \epsilon_t + e_t^{(j)}, \quad j = 1, \ldots, 6, \qquad (2.5)$$

where

$$L_t = L_{t-1} + R_{t-1} + \eta_{Lt}; \quad R_t = R_{t-1} + \eta_{R1};$$

$$\sum_{j=0}^{11} S_{t+j} = \eta_{St}, \qquad (2.6)$$

with $\{\epsilon_t\}$, $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ defined as in (2.1). The separate models defined by (2.5), (2.6) and (2.2) can be cast into a compact state-space representation with $\underset{\sim}{y}_t' = (y_t^{(1)}, \ldots, y_t^{(6)})$ as the input data, similar to the representation in Pfeffermann (1991). Following that representation, the survey errors (and in the present study also the census irregular terms) are included as part of the state vector so that there are no residual terms in the observation equation defined by (2.5). Unlike in Pfeffermann (1991), however, the transition matrix and the Variance-Covariance (V-C) matrix of the state error terms are not fixed in time since they depend on the design variances of the survey errors which, as explained in section 2.3, change over time.

The state-space representation of the model permits us to update, smooth or predict the state vectors and hence the seasonal, trend and population values at any given month t by means of the Kalman filter. Denote by $\underset{\sim}{\alpha}_t$ the state vector corresponding to month t. The state vector comprises the trend level, increment and seasonal effects, the rotation group biases and the survey errors. See Pfeffermann (1991) for details. By "updating" we mean estimation of $\underset{\sim}{\alpha}_t$ at month t based on all the data until and including month t. "Smoothing" refers to the estimation of $\underset{\sim}{\alpha}_t$ based on all the available data for all the months before and after month t. Smoothing is required for improving past estimates as, for example, when estimating the seasonal effects or when estimating changes in the population values or the trend levels. "Prediction" of state vectors corresponding to postsample months is important for policy making. Predictions within the sample period allow to assess the performance of the model, *e.g.* by comparing the forecasted panel estimates as derived from the predicted state vectors with the actual estimates. See section 3 for details. The theory of state-space models and the Kalman filter is developed in numerous publications, see Pfeffermann (1991) for the filtering and smoothing equations with references. Notice that the filtering and the smoothing equations not only yield the three sets of estimators for any given month t but also the V-C matrices of the corresponding estimation errors.

The actual application of the Kalman filter requires the estimation of the unknown model parameters and the initialization of the filter, that is, the estimation of the initial state vector $\underset{\sim}{\alpha}_0$ and the corresponding V-C matrix of the estimation errors. For a single small area, the unknown model parameters are the four variances of the error terms in the census model (2.1) and the eight autoregression coefficients and six residual variances in the panel survey error models (2.2). (The rotation group means are included in the state vectors as fixed, time invariant coefficients). In order to reduce the number of free parameters in the combined state-space model, we assume $\sigma_j^2 = \sigma^2 \times \tilde{\sigma}_j^2, j = 1, \ldots, 6$, where $\{\sigma_j^2\}$ are the residual variances in (2.2) and $\tilde{\sigma}_j^2$ are the estimates of the residual variances obtained by fitting the autoregression equations to the pseudo survey errors $e_{t,p}^{(j)}$ defined in section 2.3. This assumption reduces the number of unknown parameters from 18 to 13. (The estimates $\tilde{\sigma}_j^2$ are very close for $j = 4, 5, 6$ and have been set equal).

Assuming that the error terms in the census and survey error models have a normal distribution, the unknown model parameters can be estimated by maximization of the likelihood. See Pfeffermann and Burck (1991) for a brief description of the application of the method of scoring maximization algorithm and for the initialization of the filter. That article includes references to more rigorous discussions.

2.5 Adjustments to Account for the Use of Estimated Parameter Values

Once the unknown model parameters have been estimated, the Kalman filter equations can be applied with the true parameter values replaced by the parameter estimates. As noted in section 2.4, the Kalman filter not only produces estimates for the state vectors but also the V-C matrices of the corresponding estimation errors. A possible problem arising from the use of these V-C matrices, however, is that they ignore the extra variation implied by parameter estimation, thus resulting in underestimation of the true variances.

Formally, let $\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})$ define the estimator of $\underset{\sim}{\alpha}_t$ at month t, based on all the data available until some given month n, where $\hat{\underset{\sim}{\lambda}}$ represents the estimators of the unknown model parameters. The estimation error can be decomposed as

$$[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t] = [\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t] + [\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})], \quad (2.7)$$

which is the sum of the error if $\underset{\sim}{\lambda}$ were known plus the error due to estimation of $\underset{\sim}{\lambda}$. The two terms in the right-hand side of (2.7) are uncorrelated. A simple way to verify this property is by noting that $\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) = \mathrm{E}(\underset{\sim}{\alpha}_t \mid Y, \underset{\sim}{\lambda})$ where Y represents all the available data. By conditioning on Y and $\underset{\sim}{\lambda}$, $[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})]$ is nonstochastic whereas $\mathrm{E}\{[\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t] \mid Y, \underset{\sim}{\lambda}\} = \underset{\sim}{0}$. It follows therefore from (2.7) that

$$\begin{aligned} Q_t &= \mathrm{E}\{[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t][\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t]'\} \\ &= \mathrm{E}\{[\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t][\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t]'\} \\ &\quad + \mathrm{E}\{[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})][\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})]'\} \\ &= A_t + B_t. \end{aligned} \quad (2.8)$$

In order to estimate A_t and B_t we condition on Y and follow the approach proposed by Hamilton (1986). By this approach, realizations $\underset{\sim}{\lambda}_{(k)}, k = 1, \ldots, K$ are generated from the asymptotic normal posterior distribution of $\underset{\sim}{\lambda}$, that is, from a $N(\hat{\underset{\sim}{\lambda}}, \hat{\Lambda})$ distribution where $\hat{\underset{\sim}{\lambda}}$ is the maximum likelihood estimator of $\underset{\sim}{\lambda}$ and $\hat{\Lambda}$ is the asymptotic V-C matrix of $\hat{\underset{\sim}{\lambda}}$. (Both $\hat{\underset{\sim}{\lambda}}$ and $\hat{\Lambda}$ are obtained from the method of scoring). The Kalman filter is then applied with each of these realizations yielding estimates $\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)})$ with V-C matrices $P_t(\underset{\sim}{\lambda}_{(k)})$. The matrices A_t and B_t are estimated as

$$\hat{A}_t = \frac{1}{k} \sum_{k=1}^{K} P_t(\underset{\sim}{\lambda}_{(k)});$$

$$\hat{B}_t = \frac{1}{k} \sum_{k=1}^{K} [\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)}) - \hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})][\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)}) - \hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})]'. \quad (2.9)$$

Ansley and Kohn (1986) propose an estimator for B_t based on first order Taylor series approximation. The use of their estimator is computationally less intensive but the procedure proposed by Hamilton is somewhat more flexible in terms of the assumptions involved and it enables a better insight into the sensitivity of the Kalman filter output to errors in the parameter estimators.

2.6 Joint Modelling in Several Small Areas

The model considered so far refers to a single area. When the sample sizes in the various areas are small, more efficient estimators can often be derived by modelling in addition the cross-sectional relationships between the area population values. Clearly, the increase in efficiency resulting from such joint modelling depends on the sample sizes within the small areas and the closeness of the behaviours of the area population values over time.

The survey errors are independent between the areas so that any joint modelling of the survey estimators applies only to the census model. For modelling the unemployment rates in the four Atlantic provinces, we follow Pfeffermann and Burck (1990) and allow for nonzero contemporary correlations between corresponding error terms of the census models operating in these provinces. Thus, if $\underset{\sim}{\nu}'_{t,a} = (\epsilon_t^{(a)}, \eta_{Lt}^{(a)}, \eta_{Rt}^{(a)}, \eta_{St}^{(a)})$ denotes the vector of error terms at time t associated with the census model operating in area a, it is assumed that $C_{a,b} = \mathrm{E}(\underset{\sim}{\nu}_{ta}\, \underset{\sim}{\nu}'_{tb})$ is diagonal but with possibly non zero covariances on the main diagonal. The actual implication of this assumption is that if, for example, there is a significant increase in the trend level in one province, similar increases can be expected to occur in other provinces.

The resulting joint model holding for the four provinces (or more generally for a group of areas) can again be cast into a state-space form, see equations (2.7) and (2.8) in

Pfeffermann and Burck (1990). A major problem with the fitting of this model, however, is the joint estimation of all the unknown parameters which is computationally too intensive in terms of computer time and storage space. (The computer program written for the application of the method of scoring uses numerical first order derivatives so that each derivative requires a separate sweep through all the data. Each sweep involves the computation of the Kalman filter equations for each month included in the sample period).

To deal with this problem, we first fitted the models defined by (2.5), (2.6) and (2.2) separately for each of the provinces. We also postulated equal correlations between the corresponding error terms of the separate census models across the provinces so that

$$\phi_{a,b} = C_{a,a}^{-1/2}\, C_{a,b}\, C_{b,b}^{-1/2} = \phi \qquad 1 \leq a,b \leq 4, \qquad (2.10)$$

where $C_{a,a} = \mathrm{E}(\underset{\sim}{v}_{ta}\underset{\sim}{v}'_{ta})$. The four correlations maximizing the likelihood of the joint model were determined by a grid search procedure with the other model parameters held fixed at their previously estimated values.

The assumption of equal correlations reduces the number of unknown parameters considerably. It can be justified also by the small number of areas considered for this study implying that no other pre-imposed structure on these correlations can be safely detected. More substantively, a simple breakdown of the Labour Force by industry (Table 1 of Section 3) shows very similar relative frequencies in the four provinces suggesting a high degree of homogeneity in their economies.

2.7 Modifications to Protect Against Model Failures

The use of a model for the production of official statistics raises the question of how to protect against possible model failures. As discussed below, testing the model every time that new data becomes available is not feasible requiring instead the development of a built-in mechanism to ensure the robustness of the estimators when the model fails to hold.

For modelling the Labour Force series in small areas we employed the modification proposed by Pfeffermann and Burck (1990). By this modification, the updated state vector estimates at any given time t, are constraint to satisfy the condition

$$\sum_{a=1}^{A} w_{ta}\hat{Y}_{ta} = \sum_{a=1}^{A} w_{ta}\bar{y}_{ta} \quad t = 1, 2, \ldots, \qquad (2.11)$$

where $\hat{Y}_{ta}$ is the model based estimator of the population value Y_{ta} in area a, $\bar{y}_{ta} = 1/6\ \sum_{j=1}^{6} y_{ta}^{(j)}$ is the corresponding survey estimator and $w_{ta} = M_{ta}/M_t$ is the relative size of the Labour Force in that area so that $M_t = \sum_{a=1}^{4} M_{ta}$ and $\sum_{a=1}^{4} w_{ta} = 1$. Notice that $\sum_{a=1}^{A} w_{ta}\hat{Y}_{ta}$ and $\sum_{a=1}^{A} w_{ta}\bar{y}_t$ are correspondingly the model based estimator and the direct survey estimator of the aggregate population value in the group of areas considered. The condition 2.11 can be written alternatively as $\sum_{a=1}^{A} w_{ta}\bar{e}_{ta} = 0$ where $\bar{e}_{ta} = \sum_{j=1}^{6} e_{ta}^{(j)}/6$ is the mean survey error for state a. Pfeffermann and Burck (1990) show how to modify the Kalman filter equations so that it produces the constrained state vector estimator and its correct V-C matrix under the model (without the constraint), for every month t.

The rationale behind the modification is simple. It assumes that the total sample size in all the areas is sufficiently large and hence that the aggregate survey estimators can be trusted. This assumption in fact dictates the level of aggregation required, see below. By constraining the aggregate model based estimators to coincide with the aggregate survey estimators, the analyst ensures that any real change in the population values reflected in the survey estimators will be likewise reflected in the model based estimators. Notice that without constraining the estimators, sudden changes in the level of the series, for example, will be reflected in the model based estimators only after several months because these estimators depend not only on current data but also on past data. On the other hand, if no substantial changes occur, the model based estimators can be expected to satisfy approximately the constraints even without imposing them explicitly. Thus, the constrained estimators should perform almost as well as the unconstrainted estimators in regular time periods.

The assumption that the total sample size in all the areas is large and hence that the aggregate survey estimator is sufficiently close to the corresponding population value is critical. It guarantees (in high probability) that the modification will only occur when there are real changes in the population values and not as a result of large sampling errors. Admittedly, and as noted by one of the referees, in the application of the method to the Atlantic provinces described in section 3, the aggregate estimator is based on only four provinces so that its standard error is about 50 percent of the standard errors of the province survey estimators, depending on the province sample sizes. (The province survey estimators are independent, conditional on the corresponding population province values). Thus, if the constraints are to be used in practice, the aggregation should be carried out over a larger set of provinces or other small areas.

The following two alternative approaches have been suggested for dealing with the robustness problem:

(i) Perform a time series outlier detection as proposed for example in Chang, Tiao and Chen (1988).

(ii) Model the time series of proportions $\{\hat{\pi}_{ta} = \bar{y}_{ta}/\sum_{a=1}^{A}\bar{y}_{ta}, a = 1, \ldots, (A - 1)\}$ if these time series exhibit smoother behavior than the series $\{\bar{y}_{ta}\}$.

The detection of outliers is an important aspect of any modelling exercise but the question remaining is how to modify the population value estimates once observations (survey estimates) are detected as outliers. Notice in this respect that our main concern is with current estimates that is, the most recent available estimates. In Chang, Tiao and Chen (1988), the motivation for the outlier detections is to *remove* their effect from the observations so as to better understand the underlying structure of the series and improve the estimation of the model parameters. But if the cause of an outlier observation is a real shift in the level of the population values, this shift should not be removed but rather accounted for in the model based estimators. Harrison and Stevens (1976) propose to account for such changes by modifying the prior distribution of the state vectors, *e.g.* by increasing the variances of the state vector errors so as to allow for more rapid changes in the state vector estimators. See Morris and Pfeffermann (1984) for an example. Our approach of constraining the model based estimators to coincide with aggregate survey estimators provides a more automatic procedure that does not require timingly prior information.

The second approach suggested for dealing with the robustness problem is appealing since abrupt changes in the population values can be expected to cancel out in the ratios $\hat{\pi}_{ta}$. The main disadvantage of the use of this approach is that the model holding for the 'true' ratios π_{ta} is naturally very different from the model holding for the population values Y_t as defined by (2.1) and in particular, it no longer provides estimates for the trend and the seasonal effects which, as mentioned in the introduction, is one of the major uses of our approach. It is also not clear how to extract the estimates for the population values Y_t from the model holding for the ratios $\hat{\pi}_{ta}$, without some additional assumptions, like, for example, our assumption that the aggregate survey estimator is sufficiently close to the corresponding population value.

The use of constraints of the form (2.11) was previously considered by Battese, Harter and Fuller (1988) and by Pfeffermann and Barnard (1991) for analyzing cross-sectional surveys. Pfeffermann and Burck (1990) present empirical results illustrating the good performance of the modified estimators in abnormal time periods. See also section 3.

3. FITTING THE MODEL TO THE ATLANTIC PROVINCES, EMPIRICAL RESULTS

The model defined by (2.2), (2.5) (2.6) and (2.10) was fitted to the monthly panel estimators in the four Atlantic provinces in two stages. In the first stage the model defined by (2.2), (2.5) and (2.6) was fitted to each of the provinces separately. In the second stage, the correlations defining the matrix ϕ of (2.10) were estimated using a grid search procedure. (See section 2.6). The estimators obtained are, $\text{Diag}(\phi) = (0.5, 0.25, 0.80, 0.0)$. The data used for estimation of the model cover the years 1982-1988. Data for 1989 were used for model diagnostics by comparing the results within and outside the sample period.

3.1 Preliminary Analysis

Table 1 shows a breakdown of the Labour Force in the four provinces by industry. The figures in the table refer to March 1991. The (expected) sample sizes of the LFS are also shown. As can be seen, the percentage breakdowns in the four provinces are very similar justifying the assumption of equal correlations between the error terms of the census models across the provinces. The similarity of the percentage breakdowns suggests also possible improvements in the efficiency of the model based estimators derived from the joint model over estimators which ignore the cross-sectional correlations between the province population values.

Table 1

Labour Force by Industry in the Atlantic Provinces, March 1991

	Nova Scotia		New Brunswick		Newfoundland		Prince-Edward Island	
Sample size	4,409		3,843		2,970		1,421	
	Thousands	%	Thousands	%	Thousands	%	Thousands	%
Agriculture	7	1.7	7	2.3	0.5	0.2	6.0	9.8
Other primary industry	18	4.4	13	4.2	18.0	7.7	4.0	6.6
Manufacturing	44	10.7	37	11.9	23.0	9.9	6.0	9.8
Construction	24	5.9	21	6.8	18.0	7.7	4.0	6.6
Transp. and communication	35	8.6	30	9.6	20.0	8.6	5.0	8.3
Trade and Commerce	81	19.8	61	19.6	41.0	17.6	10.0	16.4
Finance	20	4.9	12	3.9	6.0	2.6	0.5	0.8
Services	143	35.0	107	34.4	83.0	35.6	19.0	31.1
Public Administration	36	8.8	22	7.0	23.0	9.9	6.0	9.8
Unclassified	1	0.2	1	0.3	0.5	0.2	0.5	0.8
Total	409	100.0	311	100.0	233.0	100.0	61.0	100.0

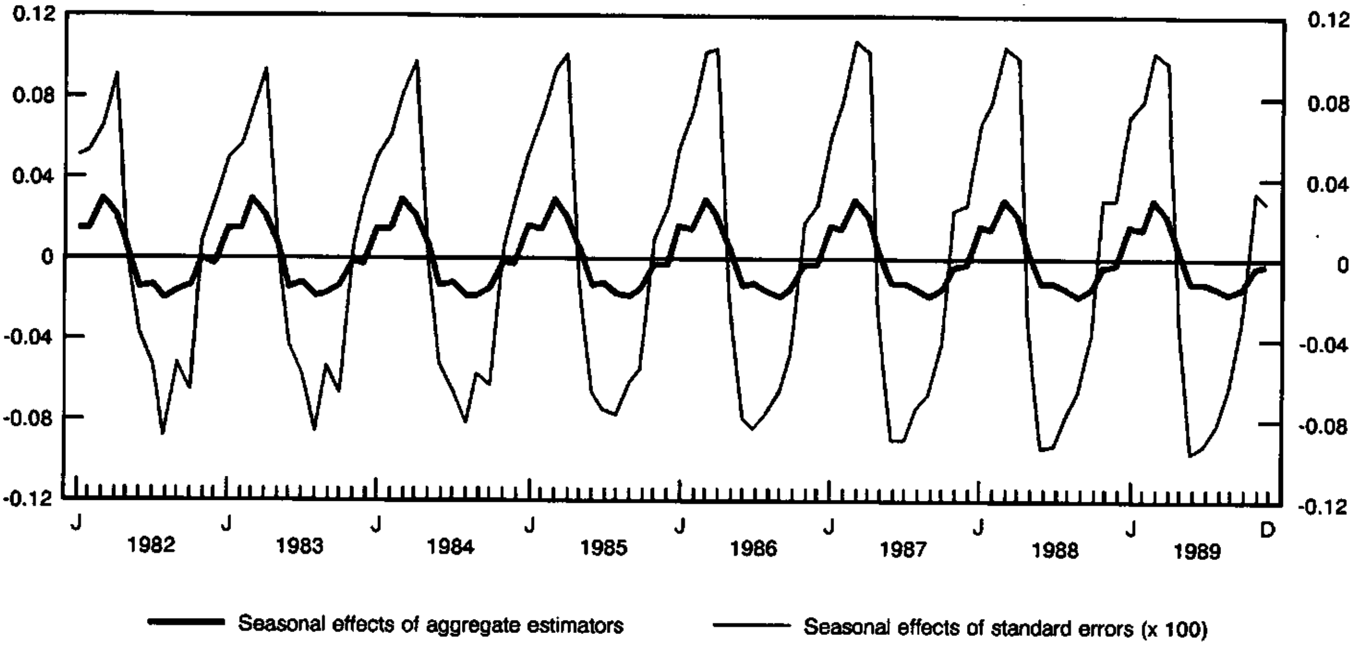

Figure 1. Seasonal Effects of Aggregate Survey Estimators and of Standard Errors of Aggregate Survey Estimators (× 100)

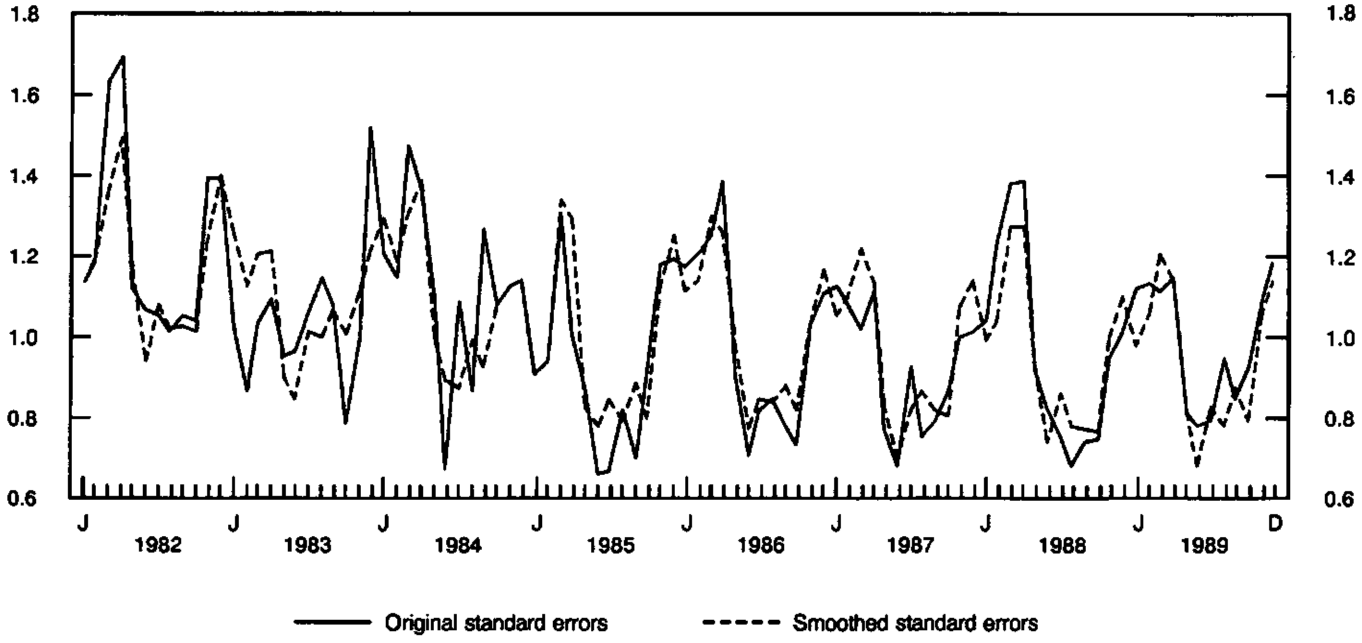

Figure 2. Original and Smoothed Standard Errors of Survey Estimators (× 100) for P.E.I. Province

Two other prior considerations mentioned in section 2.3 are that the model should account for possible rotation group effects and for changes in the variances of the survey errors over time. In order to obtain initial estimates for the rotation group effects, we averaged the pseudo survey errors, $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t), j = 1, \ldots, 6$ over all the months in the sample period. We then divided the averages by the conventional estimates of the standard errors. (The errors $e_{t,p}^{(j)}$ are correlated over time but the correlations are small because except for lags 6, 12 *etc.* the data of any given panel refer to different PSU's in the urban areas and different enumeration areas in the rural areas. See section 2.1). Notice that in the absence of rotation group effects, $\mathrm{E}(e_{t,p}^{(j)}) = 0$ for all j and t irrespective of the model postulated for the population values.

This preliminary (model free) analysis yields similar results to the results obtained under the full model, presented in Table 2 of section 3.3.

Next consider the variances of the survey errors.

Figure 1 plots the seasonal effects of the aggregate survey estimators in the four provinces along with the seasonal effects of the standard errors of these estimators (multiplied by 100). Denote as before by w_{ta} the relative labour force size in province a at time t. The aggregate survey estimator is defined as $y_t^* = \sum_{a=1}^{4} w_{ta} \bar{y}_{ta}$ (Equation 2.11). The standard error of y_t^* is $(SD^*)_t = [\sum_{a=1}^{4} w_{ta}^2 (\widehat{SD})_{ta}^2]^{1/2}$. The seasonal effects were estimated by application of the additive model of X-11 so as not to bind them to any particular model. We chose the additive model since we assume the additive decomposition for the survey estimators. (As revealed from Figure 4, the seasonal effects of the aggregate survey estimators produced by X-11 are very close to the seasonal effects obtained under the model).

Figure 1 shows that the standard errors are influenced by seasonal variations with a seasonal pattern that follows closely the seasonal pattern of the survey estimators and hence of the corresponding population values.

As discussed in section 2.3, rather than using the original estimates of the design standard errors in the models fitted to the panel survey errors we use smoothed values, thus reducing the effect of the sampling errors on the former estimators. Figure 2 plots the two sets of estimators for Prince Edward Island (P.E.I.) province which is the smallest province in the Atlantic region and hence has the smallest sample sizes. As can be seen, the effect of the smoothing is to trim the extreme raw estimates but otherwise the smoothed values behave similarly to the raw estimates. The plots for the other provinces show a similar pattern but the differences between the raw and the smoothed estimates are smaller because of the larger sample sizes in these provinces.

We conclude this section by specifying the models postulated for the seasonal effects in the four provinces. Our initial model assumed fixed variances for the error terms $\eta_{St} = \sum_{j=0}^{11} S_{t+j}, t = 1, 2, \ldots$ (see equation 2.1). The predicted errors $\hat{\eta}_{St} = \sum_{j=0}^{11} \hat{S}_{t+j}$ obtained under that model were found to decrease in absolute value as a function of time in three out of the four provinces and increase in time in the remaining province. Notice that under the model defined by (2.1), with constant variances of the state error terms, the Kalman filter converages to a steady state by which the V-C matrices of the state vector estimators and hence of $\hat{\eta}_{St}$ are constant. Thus, we modified the initial model such that $\mathrm{VAR}(\eta_{St}) = \sigma_S^2 \times g(t)$ where for the provinces of Nova Scotia, Newfoundland and P.E.I. $g(t) = t^{(-3/2)}$ whereas for New Brunswick $g(t) = t^{1/2}$.

3.2 Results

3.2.1 Rotation Group Biases

Table 2 shows the rotation group Biases (RGB) and their estimated standard errors (SE) in the four provinces as obtained under the full model defined by (2.3), (2.5), (2.6) and (2.10).

Table 2

Rotation Group Biases and Standard Errors in the Four Provinces (× 100)

Panels	Nova Scotia		New Brunswick		Newfound-land		Prince Edward Island	
	RGB	SE	RGB	SE	RGB	SE	RGB	SE
1	−0.20	0.10	−0.02	0.11	−0.47	0.13	0.32	0.17
2	0.18	0.09	0.40	0.10	0.42	0.12	0.18	0.15
3	0.32	0.08	0.24	0.09	0.47	0.12	0.31	0.15
4	0.06	0.07	0.01	0.09	0.18	0.12	0.03	0.15
5	−0.03	0.08	−0.15	0.10	−0.10	0.13	−0.25	0.16
6	−0.34	0.08	−0.50	0.11	−0.50	0.14	−0.60	0.16

The RGB behave fairly consistently across the provinces. Thus, the biases for the 3rd and 6th panel are all highly significant using the conventional t-statistic, having a positive sign for the 3rd panel and a negative sign for the 6th panel. The biases for the 4th and 5th panels have again the same sign in all the provinces and they are all non-significant.

For the 2nd panel all the biases are positive but the bias in P.E.I. is not significant. (P.E.I. is the province with the smallest sample size). It is also in P.E.I. that the sign of the bias for the 1st panel is different from the signs in the other provinces.

As discussed in section 2.3, there is more than one possible reason for the existence of RGB but the results emerging from the Table provide a strong indication that whatever the reason is, the biases found for some of the panels are real and not just the outcome of sampling errors. A drawback of the present analysis, however, is that the RGB are assumed to be fixed over time. Section 4 proposes a more flexible model.

3.2.2 Goodness of Fit

A. TESTING FOR NORMALITY

Let $I^{(j)}_{ta} = (y^{(j)}_{ta} - y^{(j)}_{ta|(t-1)})$ define the innovation when predicting the j-th panel estimator one month ahead and denote $I'_{ta} = (I^{(1)}_{ta}, \ldots, I^{(6)}_{ta}$. The use of maximum likelihood estimation in this study assumes that the vectors I_{ta} are normal deviates (see section 2.4). To test this assumption, we computed the empirical distribution of the standardized innovations $\{(SI)^{(j)}_{ta} = [I^{(j)}_{ta}/\widehat{SD}(I^{(j)}_{ta})],$ $t = (k+1), \ldots, N\}$ and compared it to the standard normal distribution using the Kolmogorov-Smirnov test statistic. This test statistic was computed for each of the six panels in the four provinces yielding P-values larger than 0.15 in 21 out of the 24 cases. (The tests were performed using PROC UNIVARIATE of the SAS package. By this procedure, if the sample size is greater than fifty as it is in our case, the data are tested against a normal distribution with mean and variance equal to the sample mean and variance). Applying the same test procedure to the standardized innovations $\{(SI)_{ta} = [I_{ta}/\widehat{SD}(I_{ta})],$ $t = (k+1), \ldots, N\}$ where $I_{ta} = [\sum^6_{j=1} I^{(j)}_{ta}/6]$ yields P-values larger than 0.15 in all the four provinces.

The estimators of the standard deviations of the innovations used for the tests are those produced by the Kalman filter, without accounting for the variance component resulting from parameter estimation (see section 2.5). The latter component is negligible even in P.E.I. which has the smallest samples sizes among the four provinces. We come back to this finding in section 3.4.

B. PREDICTION ERRORS WITH DIFFERENT PREDICTORS

Table 3 contains summary statistics comparing the behaviour of the prediction errors (innovations) in the four provinces as obtained for three different sets of estimators of the state vectors: (1) The estimators obtained under the separate models (SM) defined by (2.2), (2.5) and 2.6; (2) the estimators obtained under the joint model (JM) defined by (2.2), (2.5), (2.6) and (2.10); (3) the estimators obtained by imposing the robustness constraints (2.11) on the joint model (ROB). Below we define the summary statistics using as before the notation $I^{(j)}_{ta} = (y^{(j)}_{ta} - \hat{y}^{(j)}_{ta|(t-1)})$ for the prediction error when predicting the j-th panel estimator one month ahead.

MB_a $= \sum^N_{t=k+1} (\sum^6_{j=1} I^{(j)}_{ta}/6)/(N-k)$ – mean bias in predicting the mean survey estimator $\bar{y}_{ta} = \sum^6_{j=1} y^{(j)}_{ta}/6$.

MAB_a $= \sum^6_{j=1} | \sum^N_{t=k+1} I^{(j)}_{ta}/(N-k) | /6$ – mean absolute bias in predicting the panel estimators.

$SQRE_a$ $= \{\sum^N_{t=k+1} [1/6 \sum^6_{j=1} I^{(j)}_{ta}/\bar{y}_{ta})]^2/(N-k)\}^{1/2}$ – square root of mean square relative prediction error in predicting the mean survey estimator.

The above summary statistics are shown separately for the sample period of July 1983 – December 1988 and for the postsample period of January 1989 – December 1989. In the latter case, the data were added one data point at a time so that for predicting the survey estimator of February 1989 for example we used the data observed until January 1989 and so forth.

Table 3
Prediction Errors in the Four Provinces,
Summary Statistics (× 100)

	Nova Scotia			New Brunswick			Newfoundland			Prince Edward Island		
	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB
						7.83 - 12.88						
MB	–.11	–.07	–.06	–.12	–.09	–.06	–.25	–.18	–.08	.06	.14	.15
MAB	.12	.11	.10	.14	.12	.11	.29	.24	.20	.20	.23	.23
SQRE	5.76	5.62	5.70	5.48	5.47	5.47	7.03	6.91	6.96	9.34	9.13	9.17
						1.89 - 12.89						
MB	.14	.11	.04	.47	.47	.46	.36	.33	.17	.84	.85	.86
MAB	.32	.32	.30	.51	.51	.50	.39	.37	.29	.84	.85	.86
SQRE	6.39	6.27	6.82	6.25	6.25	6.32	5.92	5.90	5.61	9.45	9.26	9.30

The main conclusions from Table 3 are as follows:

(1) The results obtained for the three sets of predictors are in general very similar, indicating that for the data analyzed the use of the joint model improves only slightly over the use of the separate models and that there are no abrupt changes in the level of the series in the years considered.

(2) The errors when predicting the survey estimators are small both within and outside the sample period, suggesting a good fit of the model. Notice that except in P.E.I., the relative prediction errors as measured by the statistics $SQRE_a$ are all less than 7%.

(3) The biases of the prediction errors in the postsample period are larger than in the sample period with relatively large differences in New Brunswick and P.E.I. This outcome by itself could suggest some model failure in the year 1989. Inspection of the monthly panel prediction errors in the four provinces for this year, (not shown in the Table), indicates however that although the errors are in general mostly positive, the relatively large biases are mainly the result of one or two extreme errors which, with only 12 data points, has a large effect on the average summary statistics. It should be noted also that the estimated unemployment rates in the four provinces in the year 1989 are between 0.11 and 0.18 so that a prediction bias of .005 or even .009 as obtained for P.E.I. is not high. Clearly, the model can be modified to account for these biases if they persist with additional data. On the other hand, notice that the discussion above refers only to the bias of the prediction errors since the bias of the model based estimators of the concurrent population values is controlled by the robustness constraints (2.11).

In view of the very similar results obtained for the three sets of predictors considered and in order to highlight the performance of the robustness constraints, we deliberately deflated the unemployment rates in the period March 1985 to March 1987 by 33%, deflated the rates in the period April 1987 - November 1988 by 25% and inflated the rates in the period December 1988 - December 1989 by 33%. The effect of these operations is to introduce sudden drifts in the data in the months $t = 39$, $t = 64$ and $t = 84$. Figure 3 displays the aggregate, one step ahead prediction errors (APE), $I_t^a = \sum_{a=1}^{4} w_{ta} [\sum_{j=1}^{6} (y_{ta}^{(j)} - \hat{y}_{ta|(t-1)}^{(j)})/6]$ as obtained for the joint model with and without the robustness constraints, and for the separate models.

The clear conclusion from Figure 3 is that by imposing the constraints, the APE in the periods following the three months with sudden drifts are smaller than the APE

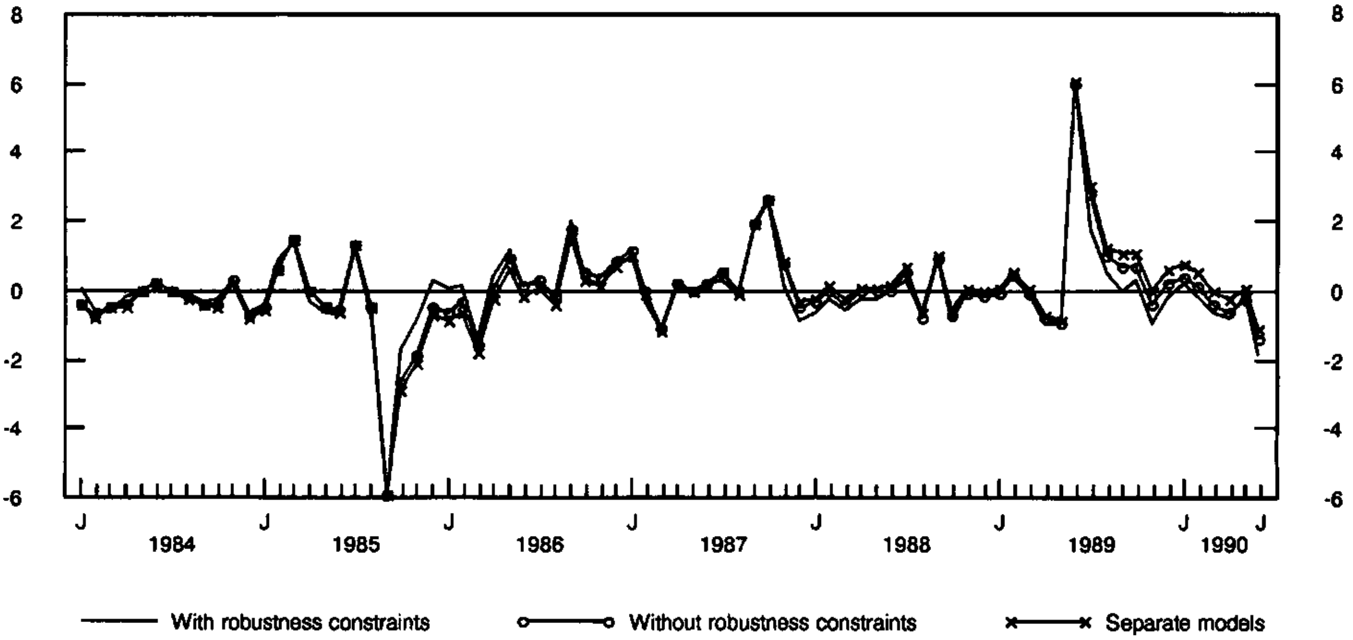

Figure 3. Aggregate One-Step Ahead Prediction Errors of the Three Sets of Predictors (× 100) for Contaminated Data

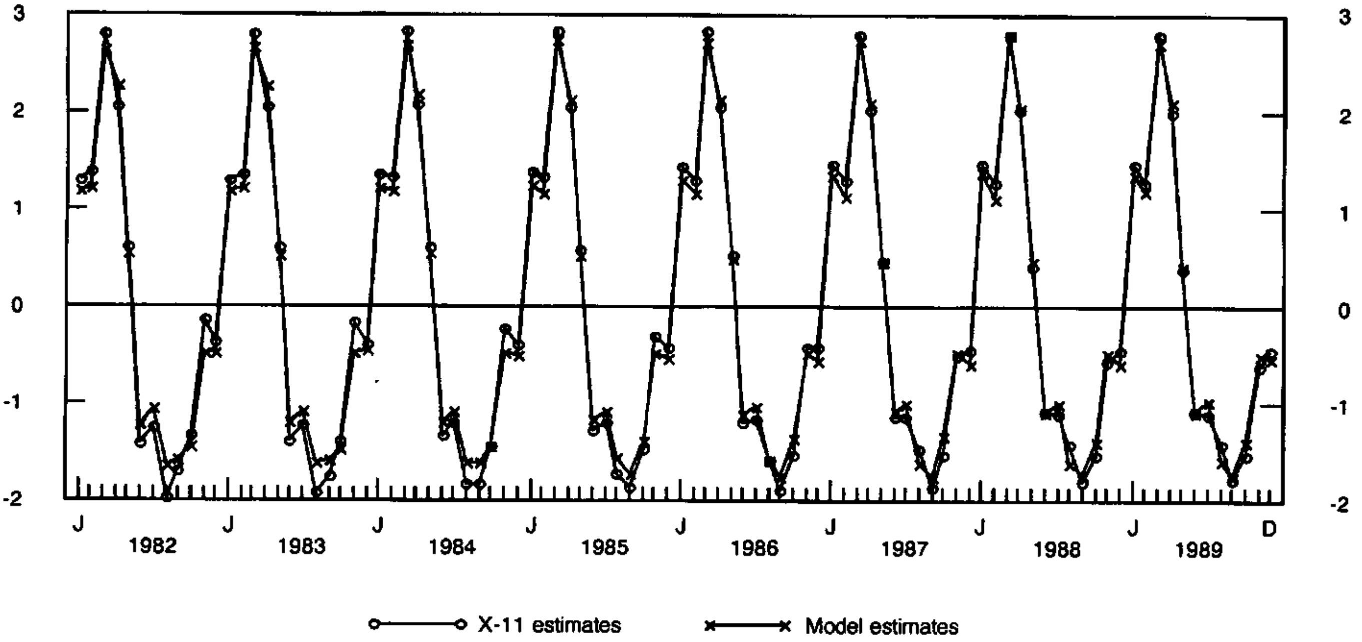

Figure 4. Weighted Averages of Seasonal Effects as Obtained by X-11 and Under the Model (× 100)

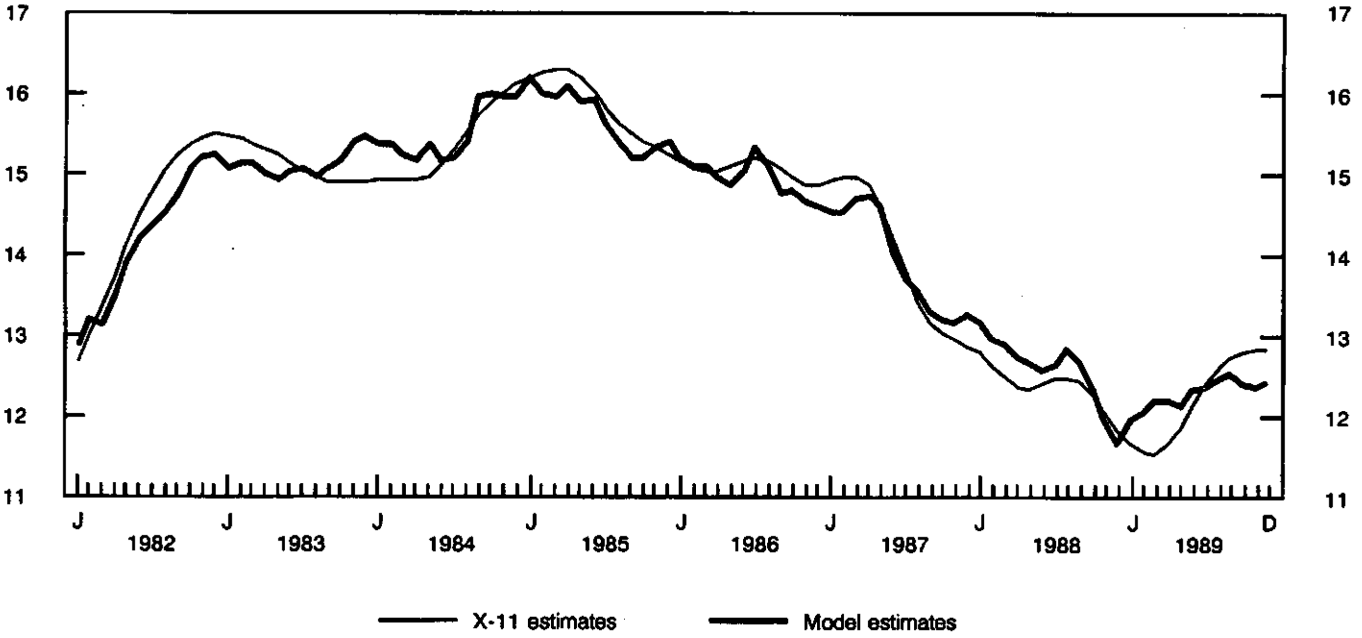

Figure 5. Weighted Averages of Trend Levels as Obtained by X-11 and Under the Model

obtained without the constraints. Thus, in March 1985 for example, $t = 39$, the APE are very large in absolute value both with and without the constraints which is obvious since the predictors use only the data until February 1985. The APE corresponding to the robust predictors return however, to their normal level much faster than the APE of the nonrobust predictors. A similar behaviour is seen to hold in the other two periods. Another notable result featured in the graph is that in the periods following the months with the sudden drifts, the joint model performs better than the separate models even without imposing the robustness constraints. Thus, by borrowing information from one province to the other, the joint model adapts itself more rapidly to the new level of the series. For more illustrations of the performance of the robustness constraints see Pfeffermann and Burck (1990).

C. COMPARISONS WITH ESTIMATORS PRODUCED BY X-11

As a final assessment of the appropriateness of the model, we compare the estimates of the seasonal effects and the trend levels as obtained under the model, with the estimates produced by the X-11 procedure (Dagum 1980). The latter is known to be less dependent on specific model assumptions. This procedure is the commonly used method for seasonal adjustment throughout the world. Figure 4 displays the average seasonal effects for the four provinces as obtained by X-11 and under the model. Figure 5 displays the corresponding trend level estimates. The averages are computed using the weights (w_{ta}) employed in previous analyses. The model based estimates shown in the two figures are the smoothed estimates which, like X-11, employ all the data in the sample period.

As can be seen, the seasonal effects produced by the two approaches are very close. The trend level estimates are also close but the X-11 trend curve is smoother than the model curve. Similar close correspondence between X-11 and the model is obtained for each of the four provinces separately, including, in particular, P.E.I. with its relatively small sample sizes.

3.3 Comparison of Design Based and Model Dependent Estimators

We mention in the introduction that one of the major reasons for wishing to model the raw survey estimators is that the model produces estimates for the population values which, at least in small areas, are more accurate (when the model holds) than the survey estimators. We computed the two sets of estimates for the four provinces and found that as expected, the estimates produced by the two approaches behave very similar but the design based estimators are less stable, having in general higher peaks and lower troughs. An important aspect when comparing the two sets of estimates is their performance in estimating year to year changes of the population values. Such comparisons are free of the obscuring effects of seasonality. Figure 6 displays the results obtained for P.E.I.. The model dependent estimates are the smoothed values of the joint model which use all the data in all the months. As can be seen, the estimates produced by the model are much more stable and vary only mildly from one month to the other compared to the design based estimates. Figure 7 displays the standard errors (S.E.) of the unemployment rates estimators in P.E.I. as computed under the design, (smoothed values, see Figure 2), and under the joint model. Also shown are the S.E. when fitting the separate model defined by (2.2), (2.5) and (2.6) and the corresponding S.E. after accounting for the use of parameter estimates instead of the unknown parameter values. See section 2.5 for details. (The latter have been computed only for the separate model to save in computing time).

There are three notable features emerging from the graphs:

(1) The S.E. of the model dependent estimators under the joint model are only mildly smaller than the S.E. obtained for the separate model but considerably smaller than the S.E. of the survey estimators.

(2) The S.E. of the model dependent estimators behave similarly to the S.E. of the survey estimators, a direct consequence of accounting for the changes in the variances of the survey errors over time in the model. See section 2.3 for details.

(3) Accounting for the use of estimated parameter values in the computation of the S.E. of the model dependent estimators has only a marginal effect on the computed S.E. Recall that P.E.I. is the province with the smallest sample sizes. The effect of accounting for the use of parameter estimates in the other provinces is even smaller.

4. SUMMARY

This article illustrates that data collected by a complex sampling design, consisting of several stages of selection with rotating panels, can be successfully modelled by a relatively simple model. The model consists of two parts: the census model holding for the population values and the survey errors model describing the time series relationship between the survey errors. The use of the model yields more accurate estimators for the population values and their components like trend and seasonality and it permits estimating the S.E. of these estimators in a rather simple way. The model equations can be modified to secure the robustness of the model-dependent estimators against possible model failures.

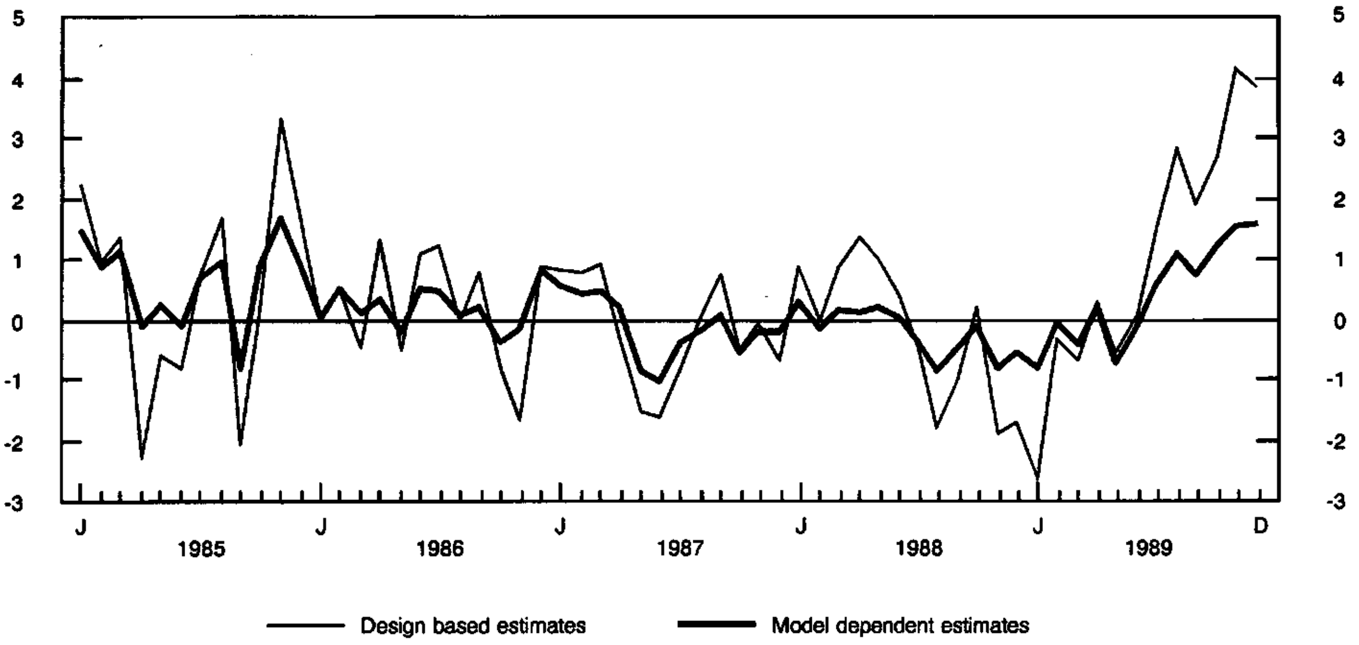

Figure 6. Year to Year Changes in Design Based and Model Dependent Estimates of P.E.I. Unemployment Rates (× 100)

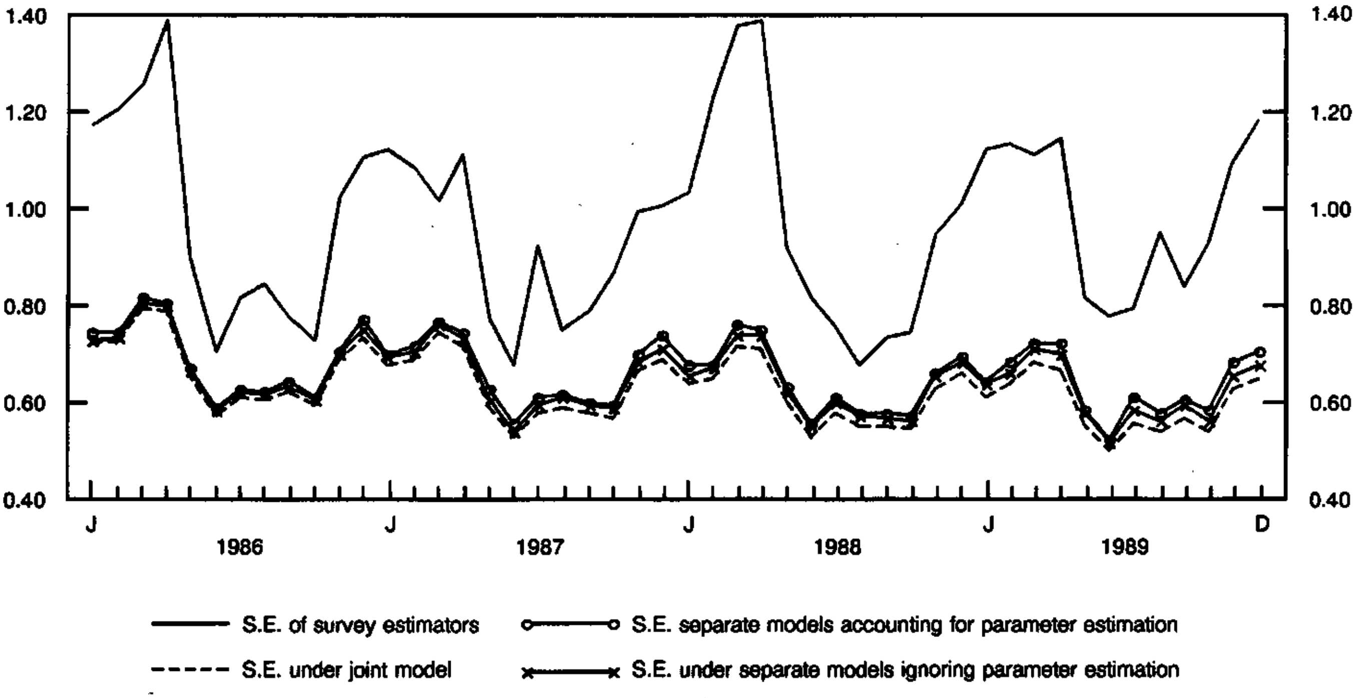

Figure 7. S.E. of Survey Estimators and of Model Dependent Estimators With and Without Accounting for Parameter Estimation (× 100) for P.E.I. Province

The model used in this article can be extended in various directions. Foremost, the model should be applied simultaneously to more provinces or other small areas to ensure that the aggregate sample estimators $\sum_{a=1}^{A} w_{ta}\bar{y}_{ta}$ are sufficiently close to the corresponding population values. See the discussion in section 2.7. Incorporating in the model an outlier detection mechanism to further assess the performance and suitability of the model is another valuable addition.

Two other extensions are to relax the assumption of constant variance for the error term ϵ_t in the census model and to let the rotation group biases to change over time. The first extension is suggested by the observation made in section 3.1 that the variances of the survey errors are subject to seasonal effects, with a seasonal pattern that is similar to the seasonal pattern of the raw estimates. Fitting the equations (2.4) in the four provinces indicates also the existence of a mild trend in the variances which again behaves similar to the trend of the raw survey estimates. Thus, the variances of the survey errors seem to depend on the magnitude of the survey estimators which suggests that the variances $\sigma_t^2 = V(\epsilon_t)$ change with the level of the population values. As a first approximation one could assume that σ_t^2 is proportional to the corresponding variance of the survey error.

Letting the rotation group biases change over time is a natural extension of the model, considering that the population values means are time dependent. Modelling the evolution of the group biases can however be problematic because of possible identifiability problems with the models holding for the trend and the seasonal effects. See the discussion in Pfeffermann (1991).

The last two extensions are important and should be explored but based on our experience with the unemployment data, we expect that they will affect the model estimators very mildly.

ACKNOWLEDGEMENT

The authors wish to thank the reviewers for their helpful comments and suggestions. Work on this study was carried out while the first author was staying at Statistics Canada under its Research Fellowship Program.

REFERENCES

ANSLEY, C.F., and KOHN, R. (1986). Predicted mean square error for state-space models with estimated parameters. *Biometrika*, 73, 467-473.

BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

CHANG, I., TIAO, G.C., and CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.

DAGUM, E.B. (1980). *The* X-11 *ARIMA Seasonal Adjustment Method*. Catalogue No. 12-564E, Statistics Canada, Ottawa, Ontario K1A 0T6.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 387-397.

HARRISON, P.J., and STEVENS, C.F. (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society*, Series B, 38, 205-247.

HARVEY, A.C. (1984). A unified view of statistical forecasting procedures (with discussion). *Journal of Forecasting*, 3, 245-275.

HARVEY, A.C., and TODD, P.H.J. (1983). Forecasting economic time series with structural and Box-Jenkins models (with discussion). *Journal of Business and Economic Statistics*, 1, 299-315.

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

MARAVALL, A. (1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics*, 3, 350-355.

MORRIS, N.D., and PFEFFERMANN, D. (1984). A Kalman filter approach to the forecasting of monthly time series affected by moving festivals. *Journal of Time Series*, 5, 255-268.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PFEFFERMANN, D., and BARNARD, C.M. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9, 73-84.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada, Ottawa, Ontario, K1A 0T6.

TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 165–172
Statistics Canada

Maximum Likelihood Estimation of Constant Multiplicative Bias Benchmarking Model with Application

IJAZ U.H. MIAN and NORMAND LANIEL[1]

ABSTRACT

The maximum likelihood estimation of a non-linear benchmarking model, proposed by Laniel and Fyfe (1989; 1990), is considered. This model takes into account the biases and sampling errors associated with the original series. Since the maximum likelihood estimators of the model parameters are not obtainable in closed forms, two iterative procedures to find the maximum likelihood estimates are discussed. The closed form expressions for the asymptotic variances and covariances of the benchmarked series, and of the fitted values are also provided. The methodology is illustrated using published Canadian retail trade data.

KEY WORDS: Autocorrelations; Bias model; Generalized least squares; Sampling errors.

1. INTRODUCTION

Benchmarking methods are very commonly used for improving sub-annual survey estimates with the help of corresponding estimates, called benchmarks, from an annual survey. The improvement generally is in terms of reductions in the biases and variances of the sub-annual estimates. For example, the monthly retail trade estimates might be improved using estimates from annual retail trade surveys. The sub-annual estimates are often biased due to coverage deficiencies in the frame. Undercoverage is caused by delay in the inclusion of new businesses and non-representation of non-employer businesses in the frame. Furthermore, the variances of the sub-annual estimates are often larger than those of the corresponding annual estimates, and the sampling covariances exist between the estimates from different time periods due to overlap of the samples. On the other hand, the annual estimates can be assumed unbiased because, in practice, their frames do not suffer much from coverage deficiencies. Detailed discussions on benchmarking can be found in Laniel and Fyfe (1989; 1990), Cholette (1987; 1988), and others.

Several procedures for benchmarking time series are available in the literature. Based on a quadratic minimization approach, Denton (1971) proposed several procedures to benchmark a single time series. Cholette (1984) proposed a modified version of Denton's order one proportional variant method where he removed the starting condition to avoid transient effects. The assumptions made by authors are very unlikely to be satisfied by most economic time series. More specifically, their models assume that the bias associated with sub-annual estimates follows a random walk and that both the sub-annual and annual data are observed without sampling errors. In general the estimates come from sample surveys and hence they are subject to sampling errors.

Hillmer and Trabelsi (1987) proposed an alternate approach to benchmarking which is based on an ARIMA model (see *e.g.*, Box and Jenkins 1976). Although this approach takes into account the sampling covariances of the sub-annual and annual estimates, the approach does not accommodate biases in the sub-annual estimates. Cholette and Dagum (1989) modified the Hillmer and Trabelsi approach by replacing the ARIMA model by an "intervention" model. This approach allows the modelling of systematic effects in the time series but still possesses the same weaknesses as found in the Hillmer and Trabelsi model (Laniel and Fyfe 1990).

In order to overcome the deficiencies mentioned above, Laniel and Fyfe (1989; 1990) proposed a non-linear benchmarking model on levels. The authors provided a complex algorithm to find the generalized least squares (GLS) estimates (and their asymptotic covariances) of the model parameters. This model takes into account the sampling covariances of the sub-annual and annual estimates, and can be used when the benchmarks come either from censuses or annual overlapping samples. This model also assumes a constant multiplicative (relative) bias associated with the sub-annual level estimates. Other constant multiplicative bias benchmarking models has been proposed by Cholette (1992) and Laniel and Mian (1991). Cholette assumes a model in which both the bias and errors are multiplicative. The author used the GLS theory to find the estimates of the model parameters after making a logarithmic transformation on the model. Laniel and Mian (1991) have provided an algorithm to find the maximum likelihood estimates of a constant multiplicative bias benchmarking model with mixed (a mixture of binding and non-binding) benchmarks. The binding benchmark here is an estimate from a census (*i.e.*, an estimate with zero variance) and the non-binding benchmark on the other hand is an estimate based on a sample. The assumption

[1] Ijaz U.H. Mian and Normand Laniel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada.

of a constant multiplicative bias will be verified in practice if the rate of frame maintenance activities is relatively stable, that is, when the proportion of frame coverage deficiencies is fairly constant over time. This assumption also implies that the covered and uncovered businesses in the frame possesses the same average period-to-period ratios with respect to the variable of interest. The nature of bias associated with sub-annual estimates may vary from one time series to another. Cholette and Dagum (1991) have proposed a benchmarking method which assumes a constant additive bias associated with the sub-annual estimates.

The purpose of this paper is to consider the maximum likelihood (ML) estimation of the parameters of Laniel and Fyfe's model and the results are based on the report of Mian and Laniel (1991). Their model is described in the next section. Two iterative processes to find the ML estimates of the model parameters are discussed in Section 3. The closed form expressions for the asymptotic covariances of the estimators of model parameters and of the fitted values are provided in Section 4. The published Canadian retail trade data collected by Statistics Canada are used to illustrate the methodology.

2. CONSTANT MULTIPLICATIVE BIAS MODEL (CMBM)

In order to meet the benchmarking requirements of the economic surveys, the following constant multiplicative bias model (CMBM) has been proposed by Laniel and Fyfe (1989; 1990). The model assumes that the biased sub-annual estimates y_t follow the relationship given by

$$y_t = \beta\theta_t + a_t, \quad t = 1, 2, \ldots, n \tag{2.1}$$

and the unbiased annual estimates z_T follow the relationship

$$z_T = \sum_{t \in T} \theta_t + b_T, \quad T = 1, 2, \ldots, m, \tag{2.2}$$

where the subscripts t and T denotes respectively the sub-annual and annual time periods, θ_t is the unknown fixed sub-annual parameter, β is an unknown constant bias parameter associated with y_t, and a_t and b_t are sampling errors associated respectively with y_t and z_T. The above model is a hybrid type (mixed) model in which bias is multiplicative but errors are additive.

Before proceeding further, let us define the column vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)'$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)'$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)'$, and $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$. The CMBM model, given by (2.1) and (2.2), can be rewritten as

$$\begin{aligned} w &= X_\beta \boldsymbol{\Theta} + u \\ &= X_{\Theta}\beta + X_D\boldsymbol{\Theta} + u, \end{aligned} \tag{2.3}$$

where

$$X_\beta = (\beta I_n : D')', \quad X_\Theta = (\boldsymbol{\Theta}' : \boldsymbol{0}')', \quad X_D = (\boldsymbol{0}' : D')',$$
$$w = (y' : z')', \qquad u = (a' : b')', \qquad D = (d_{Tt}), \tag{2.4}$$

I_n is an identity matrix of order n, $\boldsymbol{0}$ is a zero vector or matrix of an appropriate order, and d_{Tt} is an indicator function equal to 1 for $t \in T$ and to 0 otherwise. It is assumed that the sampling error vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ follow multivariate normal distributions such that $\boldsymbol{a} \sim MN(\boldsymbol{0}, V_{aa})$ and $\boldsymbol{b} \sim MN(\boldsymbol{0}, V_{bb})$. Also, in the general case, $\boldsymbol{a}$ and $\boldsymbol{b}$ are correlated, which means that $\mathrm{Cov}(\boldsymbol{a},\boldsymbol{b}) = V_{ab} = V_{ba}' \neq \boldsymbol{0}$. It is shown in the next section that the ML and GLS estimators of the $\boldsymbol{\Theta}$ and β are same for this model. Thus the assumption regarding the normality of $\boldsymbol{a}$ and $\boldsymbol{b}$ is required only to obtain the Fisher information matrix (and hence variances) of the ML estimators.

3. MAXIMUM LIKELIHOOD ESTIMATION

The log-likelihood function under CMBM can be written as

$$\ln(L) = -\frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2}\ln|V| - \frac{1}{2}Q, \tag{3.1}$$

where

$$Q = (w - X_\beta\boldsymbol{\Theta})' V^{-1} (w - X_\beta\boldsymbol{\Theta}) \tag{3.2}$$

and

$$V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}.$$

The ML estimates of the model parameters $\boldsymbol{\Theta}$ and β can be obtained, assuming V known, by maximizing the log-likelihood function (3.1) or equivalently by minimizing the quadratic term Q (3.2). For this particular model, the ML and GLS estimators of the model parameters are the same and the distinction between them will be made only if the need arises. Taking the first order partial derivatives of $\ln(L)$ with respect to $\boldsymbol{\Theta}$ and β, respectively, and then equating them to zero, we have

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \boldsymbol{\Theta}} &= X_\beta' V^{-1}(w - X_\beta\boldsymbol{\Theta}) = \boldsymbol{0}, \\ \frac{\partial \ln(L)}{\partial \beta} &= X_\Theta' V^{-1}(w - X_\beta\boldsymbol{\Theta}) = 0. \end{aligned} \tag{3.3}$$

Since $\mathrm{E}(\boldsymbol{w}) = X_\beta \boldsymbol{\Theta}$ under the model (2.3), the above equations are estimating equations in the sense of Godambe (1960) and they are information unbiased. It is interesting to note that $X_\beta' V^{-1}$ and $X_\Theta' V^{-1}$ do not depend on $\boldsymbol{w}$ so that the equations (3.3) converge to zeros and hence have consistent roots as long as $\mathrm{E}(\boldsymbol{w}) = X_\beta \boldsymbol{\Theta}$. That is, even when V in the above equations is replaced by some of its consistent estimate the equations will provide consistent estimates of the vector $\boldsymbol{\Theta}$ and β. Also note that the above equations are non-linear in the parameters to be estimated and it is not possible to obtain explicit expressions for the estimators of $\boldsymbol{\Theta}$ and β. Therefore some iterative procedure, such as the well-known Fisher-Newton-Raphson method (also called method of scores by Fisher), may be used to obtain the estimates. The elements of expected Fisher information matrix needed to implement the Fisher-Newton-Raphson method are provided in Section 4.

An alternate way to find the ML estimates of the model parameters is to solve the estimating equations (3.3) successively. By solving the first expression of (3.3), the estimate of $\boldsymbol{\Theta}$, as a function of β, is given by

$$\hat{\boldsymbol{\Theta}} \equiv \hat{\boldsymbol{\Theta}}(\beta) = (X_\beta' V^{-1} X_\beta)^{-1} X_\beta' V^{-1} \boldsymbol{w}. \qquad (3.4)$$

Similarly, by solving the second expression of (3.3), the estimator of β, as a function of $\boldsymbol{\Theta}$, is given by

$$\hat{\beta} \equiv \hat{\beta}(\boldsymbol{\Theta}) = \left[\boldsymbol{\Theta}' V_{aa.b}^{-1}\left(y - V_{ab} V_{bb}^{-1}(z - \boldsymbol{D\Theta})\right)\right] \Big/ \left[\boldsymbol{\Theta}' V_{aa.b}^{-1} \boldsymbol{\Theta}\right], \qquad (3.5)$$

where

$$V_{aa.b} = V_{aa} - V_{ab} V_{bb}^{-1} V_{ba}.$$

The ML estimates of $\boldsymbol{\Theta}$ and β can be obtained by successively calculating equations (3.4) and (3.5) until convergence. This procedure has an advantage over the Fisher-Newton-Raphson method as it is easy to implement. However, for this kind of algorithm, the convergence is usually very slow. We will compare these two methods in Section 6 to check the speed of their convergence.

Once the ML estimates of the model parameters are obtained, one can find the fitted sub-annual values $\hat{y} = \hat{\beta}\hat{\boldsymbol{\Theta}}$ and the fitted annual values $\hat{z} = \boldsymbol{D}\hat{\boldsymbol{\Theta}}$.

Initial Guess for $\boldsymbol{\Theta}$ and β

In order to obtain an initial guess for β, say $\hat{\beta}_0$, let us rewrite the model (2.3) as

$$\boldsymbol{w}^* = X_\Theta^* \beta + \boldsymbol{u}^*,$$

where $\boldsymbol{w}^* = ((\boldsymbol{D}y)':(z - \boldsymbol{D\Theta})')'$, $X_\Theta^* = \left((\boldsymbol{D\Theta})':0'\right)'$ and $\boldsymbol{u}^* = \left((\boldsymbol{D}a)':b'\right)'$. Thus the ML estimate of β is given by

$$\hat{\beta} = \left[X_\Theta^{*\prime} (V^*)^{-1} \boldsymbol{w}^*\right] \Big/ \left[X_\Theta^{*\prime} (V^*)^{-1} X_\Theta^*\right], \qquad (3.6)$$

where

$$V^* = \mathrm{Cov}(\boldsymbol{u}^*) = \begin{pmatrix} \boldsymbol{D} V_{aa} \boldsymbol{D}' & \boldsymbol{D} V_{ab} \\ V_{ba} \boldsymbol{D}' & V_{bb} \end{pmatrix}.$$

Using the fact that $\mathrm{E}(z) = \boldsymbol{D\Theta}$, and replacing $\boldsymbol{D\Theta}$ by z in (3.6), an initial guess for β may be taken as

$$\hat{\beta}_0 = \left[\begin{pmatrix} z \\ 0 \end{pmatrix}' (V^*)^{-1} \boldsymbol{w}^*\right] \Big/ \left[\begin{pmatrix} z \\ 0 \end{pmatrix}' (V^*)^{-1} \begin{pmatrix} z \\ 0 \end{pmatrix}\right]$$

$$= \left[z' \left(\boldsymbol{D} V_{aa.b} \boldsymbol{D}'\right)^{-1} \boldsymbol{D}y\right] \Big/ \left[z' \left(\boldsymbol{D} V_{aa.b} \boldsymbol{D}'\right)^{-1} z\right]. \qquad (3.7)$$

The initial estimate of $\boldsymbol{\Theta}$ can be obtained from (3.4) by replacing β by $\hat{\beta}_0$.

4. COVARIANCES OF THE ESTIMATORS

In this section, we derive the expressions for the asymptotic covariances of the ML estimators of CMBM parameters by inverting the Fisher information matrix, say $\boldsymbol{\Omega}$. The asymptotic covariances of the fitted sub-annual and annual values are provided by using the delta method. First, let us consider the derivation of the covariances of the ML estimators of $\boldsymbol{\Theta}$ and β. The elements of $\boldsymbol{\Omega}$ (*i.e.*, the negative expectations of the second order partial derivatives of $\ln(L)$) are given by

$$\Omega_{11} = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \boldsymbol{\Theta}\, \partial \boldsymbol{\Theta}'}\right] = X_\beta' V^{-1} X_\beta,$$

$$\Omega_{22} = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \beta^2}\right] = \boldsymbol{\Theta}' V_{aa.b}^{-1} \boldsymbol{\Theta}$$

and

$$\Omega_{12} = \Omega_{21}' = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \boldsymbol{\Theta}\, \partial \beta}\right] = X_\beta' V^{-1} X_\Theta.$$

Therefore, the Fisher information matrix of order $(n + 1) \times (n + 1)$ is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \qquad (4.1)$$

Inverting Ω by using the algebra of partitioned matrices we have

$$\begin{aligned} \operatorname{Cov}(\hat{\boldsymbol{\Theta}}) &= \Omega_{11.2}^{-1}, \\ \operatorname{Var}(\hat{\beta}) &= \Omega_{22.1}^{-1}, \\ \operatorname{Cov}(\hat{\boldsymbol{\Theta}},\hat{\beta}) &= -\,\Omega_{11.2}^{-1}\Omega_{12}\Omega_{22}^{-1} \\ &= -\,\Omega_{11}^{-1}\Omega_{12}\Omega_{22.1}^{-1}, \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \Omega_{11.2} &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, \\ \Omega_{22.1} &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}. \end{aligned} \tag{4.3}$$

Once the covariance matrix Ω^{-1} is available, the asymptotic covariances of the sub-annual fitted values $\hat{y}$ can be obtained by using the delta method (see *e.g.*, Rao 1973). Let Δ be the matrix of first order partial derivatives of y with respect to the elements of $(\boldsymbol{\Theta}':\beta)'$. Clearly, the $n \times (n+1)$ matrix is $\Delta = (\beta I_n : \boldsymbol{\Theta})$. Now, by using the delta method, the asymptotic covariance matrix of $\hat{y}$ is given by

$$\operatorname{Cov}(\hat{y}) = \Delta\Omega^{-1}\Delta'. \tag{4.4}$$

Furthermore, the covariance matrix of the annual fitted values $\hat{z}$, from the standard multivariate normal theory, is given by

$$\operatorname{Cov}(\hat{z}) = D\Omega_{11.2}^{-1}D', \tag{4.5}$$

where D and $\Omega_{11.2}$ are as defined by (2.4) and (4.3), respectively.

5. MAXIMUM LIKELIHOOD ESTIMATION WHEN $V_{ab} = \mathbf{0}$

In this section we consider the ML estimation of the model parameters for the special case when the error vectors a and b are uncorrelated (*i.e.*, $\operatorname{Cov}(a,b) = V_{ab} = V_{ba}' = \mathbf{0}$). Usually this is the case in sample surveys when annual and sub-annual samples are drawn independently from each other. Reduction in the results of Sections 3 and 4 can be seen by substituting $V_{ab} = V_{ba}' = \mathbf{0}$ in the equations. As an example, for this special case, the ML estimators of $\boldsymbol{\Theta}$ and β, given by (3.4) and (3.5), reduce to

$$\hat{\boldsymbol{\Theta}}^* \equiv \hat{\boldsymbol{\Theta}}^*(\beta) = \left(\beta^2 V_{aa}^{-1} + D'V_{bb}^{-1}D\right)^{-1} \left(\beta V_{aa}^{-1}y + D'V_{bb}^{-1}z\right)$$

and

$$\hat{\beta}^* \equiv \hat{\beta}^*(\boldsymbol{\Theta}) = \left[\boldsymbol{\Theta}'V_{aa}^{-1}y\right]/\left[\boldsymbol{\Theta}'V_{aa}^{-1}\boldsymbol{\Theta}\right],$$

respectively. These equations must be solved successively to obtain the required estimates.

Similarly, the elements of the Fisher information matrix reduce to

$$\begin{aligned} \Omega_{11}^* &= \beta^2 V_{aa}^{-1} + D'V_{bb}^{-1}D, \\ \Omega_{22}^* &= \boldsymbol{\Theta}'V_{aa}^{-1}\boldsymbol{\Theta}, \\ \Omega_{12}^* &= \Omega_{21}^{*\prime} = \beta V_{aa}^{-1}\boldsymbol{\Theta}. \end{aligned}$$

6. AN APPLICATION

Here we present an example using published Canadian retail trade data which results from monthly and annual retail trade surveys conducted by Statistics Canada. The monthly retail trade estimates and their coefficients of variation (CV) are available from the Statistics Canada publication "Retail Trade" (Catalogue No. 63-005 Monthly). There are two types of monthly retail trade estimates, namely preliminary and revised estimates. We use the revised but seasonally unadjusted (raw) estimates for this example. Since the CVs of the revised estimates are not available, the CVs of the preliminary estimates are used to approximate the variances of the revised monthly estimates. The data for the period January 1985 to December 1988 are used in this example. Another difficulty was to find the autocorrelations for monthly retail trade estimates. Based on some monthly retail trade data, Hidiroglou and Giroux (1986) provided the estimates of autocorrelations at lags 1, 3, 6, 9 and 12 for three different kinds of stratum in several provinces of Canada. As an approximation to the autocorrelations of monthly retail trade estimates, the averages of their estimates of autocorrelations for the strata in the Province of Ontario and Standard Industrial Classification Code 60 (Foods, Beverages, and Drug industries) are used. The approximate (averaged) autocorrelations, say $\rho(k)$, are given in Table 1.

Table 1

Approximate Autocorrelations $\rho(k)$ for the Monthly Retail Trade Estimates

Lag k	1	3	6	9	12
$\rho(k)$	0.970	0.940	0.918	0.914	0.962

The method of ordinary least squares and an algorithm of McLeod (1975) for the derivation of theoretical autocorrelations for autoregressive moving-average time series was used to revise the observed autocorrelations. An ARMA $(1,0)(1,0)_{12}$ seasonal multiplicative model was fitted on the five observed autocorrelations by minimizing the sum of squared differences between the observed and theoretical autocorrelations. Then the estimated model parameters and the above mentioned algorithm of McLeod were used to calculate the autocorrelations for all other lags of interest. Given that the ARMA model is correct for theoretical autocorrelations, this approach provides a consistent estimate of the autocorrelation function. These final (revised) approximate autocorrelations for up to 47 lags are given in Table 2 and were used to approximate the covariances for monthly retail trade estimates via multiplication with the standard deviations.

Table 2

Revised Approximate Autocorrelations $\rho^*(k)$ for the Monthly Retail Trade Estimates for up to 47 Lags

Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$
0	1.0000	12	0.9602	24	0.8896	36	0.8100
1	0.9758	13	0.9345	25	0.8647	37	0.7869
2	0.9555	14	0.9126	26	0.8433	38	0.7669
3	0.9391	15	0.8943	27	0.8253	39	0.7501
4	0.9266	16	0.8798	28	0.8107	40	0.7363
5	0.9177	17	0.8687	29	0.7994	41	0.7254
6	0.9126	18	0.8612	30	0.7913	42	0.7176
7	0.9113	19	0.8572	31	0.7864	43	0.7126
8	0.9136	20	0.8567	32	0.7843	44	0.7106
9	0.9196	21	0.8595	33	0.7862	45	0.7114
10	0.9293	22	0.8661	34	0.7909	46	0.7151
11	0.9429	23	0.8760	35	0.7989	47	0.7217

At the time this study was performed, the annual retail trade estimates were only available for years 1985 through 1988. These estimates are available from Statistics Canada publication "Annual Retail Trade" (Catalogue No. 63-223 Annual). The variances of annual retail trade estimates are not available from the literature and have been computed from the actual survey data. The covariances between monthly and annual estimates are zero because the samples of monthly and annual retail trade surveys were drawn independently from each other. The annual retail trade estimates are from dependent samples, thus their covariances are non-zero. But the estimates of covariances are not readily available via regular survey processing and a study would be required to obtain them. Consequently, for the purpose of this example, we assumed that the covariances between annual retail trade estimates are zero.

An interesting question was raised by one of the referees. He asked what will happen when the variances and covariances of survey estimates are not known. This is a difficult problem and cannot be answered so easily. However the model presented assumes these variances and covariances are known. In general, the estimating equations used to find the maximum likelihood estimates need only the consistent estimates of variances and covariances. It is a common practice in benchmarking problems to estimate these variances and covariances from survey data since the theoretical values are never known (see, *e.g.*, Hillmer and Trabelsi 1987).

The computations required for this example are performed by an algorithm written in the GAUSS programming language for micro computers. The initial estimate of β for the iterative process, obtained form (3.7), is given by $\hat{\beta}_0 = 0.9162$. The initial estimate of the parameter vector $\boldsymbol{\Theta}$ is obtained from (3.4), after replacing β by $\hat{\beta}_0$. Both the Fisher-Newton-Raphson and successive iteration methods, as discussed in Section 3, are used to find the ML estimates of the model parameters. The final ML estimate of β is found to be very close to the initial estimate and is given by $\hat{\beta} = 0.9016$ with CV $= 0.0065$. It is interesting to note that the Fisher-Newton-Raphson method converged very quickly to a final solution for this example. In fact it converged in only 6 iterations (about 1 minute) for a ten digit precision whereas the successive calculations method converged, with the same precision, in over 500 iterations (over 45 minutes), on a 386DX-25Mhz personal computer. However, as they should, both methods converged to the same final solution. The covariance matrix of the estimated vector $(\hat{\boldsymbol{\Theta}}':\hat{\beta})'$ is obtained by inverting the Fisher information matrix $\boldsymbol{\Omega}$, given by (4.1), after replacing parameters by their ML estimates. The original series of the monthly retail trade estimates and the benchmarked series of the ML estimates along with their CVs are given in Table 3. The fitted sub-annual series along with their CVs are also given in this table (last two columns). The original and benchmarked series are also plotted in Figure 1. The results show that the original behaviour of the series is not disturbed by benchmarking and a very large reduction in the CVs of sub-annual estimates is achieved. The original series of the annual retail trade estimates and fitted annual values along with their CVs are given in Table 4. The variances of the fitted values in Tables 3 and 4 are obtained by using expressions (4.4) and (4.5), respectively, after replacing parameters by their ML estimates. The results of fitted values also show a large reduction in the CV's of the original estimates. That is, the reliability of the monthly and annual series are increased by benchmarking.

Table 3

Monthly Retail Trade Estimates, ML Estimates of the Θ_t's and Fitted Values (all in millions of dollars) Along with their CV's

Year	Month	y_t*	CV(y_t)*	$\hat{\Theta}_t$	CV($\hat{\Theta}_t$)	$\hat{y}_t$	CV($\hat{y}_t$)
1985	1	8,689.668	0.008	9,686.630	0.00210	8,733.384	0.00667
	2	8,390.380	0.008	9,350.078	0.00210	8,429.951	0.00665
	3	10,107.485	0.006	11,248.048	0.00233	10,141.146	0.00496
	4	10,541.145	0.008	11,741.785	0.00200	10,586.294	0.00656
	5	11,763.659	0.007	13,094.151	0.00198	11,805.576	0.00570
	6	11,067.487	0.008	12,321.326	0.00189	11,108.803	0.00647
	7	10,810.755	0.008	12,029.467	0.00184	10,845.666	0.00643
	8	11,289.656	0.009	12,554.808	0.00206	11,319.309	0.00726
	9	10,336.540	0.009	11,484.216	0.00205	10,354.073	0.00728
	10	11,213.751	0.010	12,447.696	0.00256	11,222.737	0.00809
	11	11,935.495	0.010	13,234.412	0.00258	11,932.034	0.00808
	12	13,300.288	0.008	14,734.891	0.00188	13,284.853	0.00643
1986	1	9,753.373	0.009	10,794.009	0.00221	9,731.787	0.00716
	2	9,249.279	0.009	10,227.777	0.00224	9,221.277	0.00709
	3	10,609.952	0.008	11,729.293	0.00207	10,575.031	0.00622
	4	11,637.936	0.008	12,860.626	0.00206	11,595.032	0.00614
	5	12,695.108	0.008	14,024.139	0.00205	12,644.046	0.00605
	6	11,826.254	0.008	13,059.556	0.00202	11,774.385	0.00598
	7	11,940.908	0.010	13,164.500	0.00233	11,869.002	0.00740
	8	11,866.547	0.010	13,070.205	0.00232	11,783.987	0.00743
	9	11,540.397	0.009	12,712.283	0.00202	11,461.287	0.00670
	10	12,208.845	0.010	13,430.932	0.00235	12,109.215	0.00747
	11	12,201.498	0.010	13,418.219	0.00240	12,097.753	0.00747
	12	14,479.170	0.009	15,933.951	0.00215	14,365.916	0.00670
1987	1	10,271.723	0.012	11,276.676	0.00357	10,166.956	0.00891
	2	9,951.105	0.010	10,945.319	0.00261	9,868.208	0.00737
	3	11,492.162	0.008	12,663.849	0.00230	11,417.620	0.00584
	4	12,867.443	0.009	14,172.605	0.00235	12,777.901	0.00652
	5	13,508.434	0.012	14,850.145	0.00343	13,388.765	0.00862
	6	13,608.274	0.011	14,973.985	0.00287	13,500.418	0.00786
	7	13,278.474	0.023	14,483.340	0.01066	13,058.057	0.00165
	8	12,728.196	0.008	14,028.998	0.00227	12,648.426	0.00577
	9	12,616.239	0.009	13,888.982	0.00233	12,522.188	0.00659
	10	13,760.829	0.008	15,156.409	0.00227	13,664.890	0.00592
	11	13,380.142	0.008	14,733.240	0.00227	13,283.365	0.00597
	12	16,269.757	0.007	17,928.148	0.00241	16,163.867	0.00525
1988	1	11,134.013	0.010	12,234.529	0.00274	11,030.548	0.00753
	2	10,959.374	0.010	12,042.761	0.00276	10,857.651	0.00754
	3	13,177.788	0.008	14,508.565	0.00233	13,080.800	0.00602
	4	13,666.311	0.009	15,035.737	0.00243	13,556.094	0.00676
	5	14,267.530	0.006	15,742.039	0.00379	14,192.890	0.00448
	6	14,432.944	0.009	15,884.130	0.00240	14,320.997	0.00673
	7	13,960.825	0.009	15,363.957	0.00240	13,852.014	0.00673
	8	13,691.315	0.008	15,073.691	0.00233	13,590.312	0.00606
	9	13,773.109	0.008	15,159.075	0.00235	13,667.294	0.00613
	10	13,900.743	0.009	15,279.950	0.00255	13,776.282	0.00696
	11	14,453.461	0.009	15,884.279	0.00260	14,321.132	0.00700
	12	17,772.990	0.009	19,529.791	0.00267	17,607.895	0.00702

***Source**: Statistics Canada publication "Retail Trade" (Catalogue No. 63-005 Monthly).

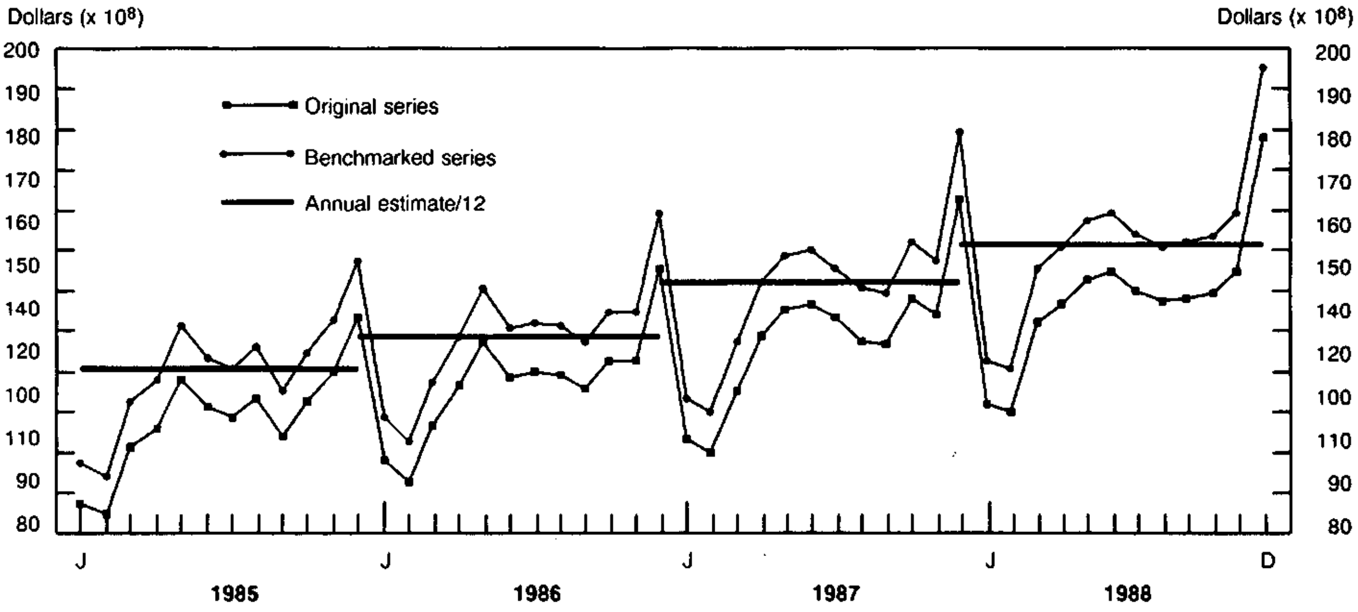

Figure 1. Original and Benchmarked Series of Monthly Retail Trade Estimates for All Stores in Canada

Table 4

Annual Retail Trade Estimates and Annual Fitted Values (in millions of dollars) Along with their CV's

Year	z_T^*	CV(z_T)	$\hat{z}_T$	CV$(\hat{z}_T)$
1985	143,965.400	0.00033	143,927.507	0.00032
1986	154,377.100	0.00031	154,425.491	0.00030
1987	169,944.600	0.00193	169,101.697	0.00128
1988	181,594.000	0.00137	181,738.512	0.00127

***Source:** Statistics Canada publication "Annual Retail Trade" (Catalogue No. 63-223 Annual).

7. CONCLUSIONS

The non-linear model discussed here seems to be very appropriate for benchmarking an economic time series from large sample surveys. The proposed iterative procedures to find the maximum likelihood estimates of the model parameters are very simple to implement in practice. However, the convergence of the successive calculation method is very slow in comparison to the Fisher-Newton-Raphson method. The closed form expressions for the covariances of the ML estimators are provided. These estimates and their covariances may be used to make inferences regarding model parameters. Furthermore, expressions for the fitted sub-annual and annual values along with their asymptotic covariances are also provided. The methodology presented in this article seems to provide a good fit to the Canadian retail trade data. However, the goodness of fit tests for this and other benchmarking models need to be developed.

ACKNOWLEDGEMENTS

We acknowledge the valuable suggestions made by the editor, associate editor and two referees which had greatly improved the quality of this paper. We are also thankful to Drs. J. Gambino and M. Kovačević of Statistics Canada for their comments on this research.

REFERENCES

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

CHOLETTE, P.A. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology,* 10, 35-49.

CHOLETTE, P.A. (1987). Benchmarking and interpolation of time series. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A. (1988). Benchmarking systems of socio-economic time series. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A. (1992). Users' manual of programmes BENCH and CALEND to benchmark, interpolate and Calendarize time series data on micro computers. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A., and DAGUM, E.B. (1989). Benchmarking socio-economic time series data: a unified approach. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A., and DAGUM, E.B. (1991). Benchmarking time series with autocorrelated sampling errors. Methodology Branch Working Paper, Statistics Canada.

DENTON, F.T. (1971). Adjustment on monthly or quarterly series to annual totals: An approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 99-102.

GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 13, 1208-1211.

HIDIROGLOU, M.A., and GIROUX, S. (1986). Composite estimation for the Retail Trade Survey. Methodology Branch Working Paper, Statistics Canada.

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

LANIEL, N., and FYFE, K. (1989). Benchmarking of economic time series. Methodology Branch Working Paper, Statistics Canada.

LANIEL, N., and FYFE, K. (1990). Benchmarking of economic time series. *Survey Methodology*, 16, 271-277.

LANIEL, N., and MIAN, I.U.H. (1991). Maximum likelihood estimation for the constant bias model with mixed benchmarks. Methodology Branch Working Paper, Statistics Canada.

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

MIAN, I.U.H., and LANIEL, N. (1991). Maximum likelihood estimation for the constant bias benchmarking model. Methodology Branch Working Paper, Statistics Canada.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd Ed.). New York: John Wiley.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 173–182
Statistics Canada

Optimum Two-Stage Sample Design for Ratio Estimators: Application to Quality Control – 1990 French Census

JEAN-CLAUDE DEVILLE[1]

ABSTRACT

This study is based on the use of superpopulation models to anticipate, before data collection, the variance of a measure by ratio sampling. The method, based on models that are both simple and fairly realistic, produces expressions of varying complexity and then optimizes them, in some cases rigorously, in others approximately. The solution to the final problem discussed points up a rarely considered factor in sample design optimization: the cost related to collecting individual information.

KEY WORDS: Census quality control; Superpopulation model; Two-stage sample design optimization; Multiple objective survey.

1. INTRODUCTION

The survey method used for quality control of French census data pointed up a number of new and interesting problems, three of which are dealt with in this paper. After discussing them in general terms, we describe their specific application to the census.

In all cases, the problem is one of optimizing a two-stage survey in which the primary units are census collection districts. Units are selected using an index k that varies in a population U of districts and is, in concrete terms, a processing unit of the census forms collected.

The first problem is that of estimating the frequency of a characteristic in the population of forms (the fact of containing an error). Keeping in mind the accuracy defined for this estimate, an attempt is made to minimize survey cost with a cost function in the form

$$C_T = mC_o + nC_1, \tag{1.1}$$

where m is the number of primary units (districts) sampled, C_o the cost of processing one PU, n the number of final units (forms) sampled and C_1 the cost of processing one final unit. The problem is fairly common when a mean is to be estimated (see for example W. Cochran (1977)). Our solution is more complete as it takes into account the great variability in primary unit size.

The second, more unique, problem is also more significant. The final population (*i.e.* the forms) is made up of G separate groups ($g = 1$ to G). We are looking for an estimate of the frequency of occurrence of a characteristic in each group, with an accuracy defined for each one. The constraint resides in the fact that, because the primary units are common to all groups, sampling within one PU affects all groups.

The objective is to minimize survey cost, which is expressed as

$$C_T = mC_o + \sum_{g=1}^{G} n_g C_g, \tag{1.2}$$

where n_g is the total number of final units in group g and C_g the cost of processing one final unit in group g. In practice the groups are made up of the different types of census forms.

The third problem is related to coding control. We do have an *a priori* measure of the difficulty of coding each form. Formally, therefore, we have, at the level of each individual i in the population, a quantitative variable X_i, such that the probability (within a meaning to be defined) of the individual having the characteristic to be measured is approximately proportional to X_i. We are seeking to use this information to minimize the cost of control (measurement of the frequency of the "coding error" characteristic) subject to a defined survey accuracy.

In each case, plausible and simple superpopulation models allow us to evaluate the anticipated variance of the survey. In a manner of speaking, this is an almost standard illustration of model assisted survey sampling as described in Särndal, Swensson, Wretman (1992).

2. OPTIMUM ESTIMATE OF THE PROPORTION OF RECORDS CONTAINING ERRORS TWO-STAGE SAMPLE DESIGN

Each primary unit k (district) has a known number N_k of individuals (forms). Of this number, D_k display the characteristic of interest (*i.e.* contain an error). The aim is to estimate:

[1] Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Études Économiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

$$P = \sum_U D_k \Big/ \sum_U N_k .$$

The survey is done by drawing a sample s of primary units (PU), with π_k, the probability of inclusion in the first order and $\pi_{k\ell}$ in the second order, to be determined. Subsequently, if primary unit k is drawn in s, n_k individuals drawn by simple random sampling without replacement are checked; d_k denotes the number of forms containing errors that will be found.

Estimator $\hat{P}_k$ of $P_k = D_k/N_k$ is expressed $\hat{P}_k = d_k/n_k$ and $\hat{D}_k = N_k\hat{P}_k$ gives an unbiased estimate of D_k. The estimator of P is expressed

$$\hat{P} = \frac{\displaystyle\sum_s \frac{\hat{D}_k}{\pi_k}}{\displaystyle\sum_s \frac{\hat{N}_k}{\pi_k}}. \tag{2.1}$$

This is the ratio of the unbiased estimators of D and N, the total number of forms. Although this number is known, estimator (4.1) is obviously more accurate than $1/N \sum_s \hat{D}_k/\pi_k$.

We have

$$\operatorname{Var}(\hat{P}) = E\operatorname{Var}(\hat{P} \mid s) + \operatorname{Var} E(\hat{P} \mid s). \tag{2.2}$$

Now

$$\operatorname{Var}(\hat{P} \mid s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1 - P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right)$$

where $$\hat{N} = \sum_s \frac{N_k}{\pi_k}.$$

Hence

$$E\operatorname{Var}(\hat{P} \mid s) \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1 - P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right). \tag{2.3}$$

Furthermore,

$$E(\hat{P} \mid s) = \frac{\displaystyle\sum_s \frac{D_k}{\pi_k}}{\displaystyle\sum_s \frac{N_k}{\pi_k}}.$$

The variance of this value is obtained by linearization following introduction of variable $Z_k = D_k - PN_k = N_k(P_k - P)$.

We obtain

$$\operatorname{Var} E(\hat{P} \mid s) \simeq N^{-2} \operatorname{Var}\left(\sum_s \frac{z_k}{\pi_k}\right).$$

Taking into account that $\sum_U Z_k = 0$:

$$\operatorname{Var}E(\hat{P} \mid s) = N^{-2}\left(\sum_k \frac{Z_k^2}{\pi_k} + \sum\sum_{k \neq l} \frac{Z_k Z_\ell}{\pi_k \pi_\ell}\pi_{k\ell}\right). \tag{2.4}$$

The sum of (2.3) and (2.4) gives us the variance of estimator (2.1).

2.1 Introduction of a Model

Not only is the variance of $\hat{P}$ difficult to manipulate, it contains unknown parameters. The problem may be circumvented by formulating the hypotheses required to produce a superpopulation model. It is assumed below that the parameters of this model may be estimated from the results of a preliminary test covering a very small portion of the population. In the model, expectation is denoted by E_ξ (variance by Var_ξ) and all the random variables are assumed independent of the sampling process.

The model has the following specifications:

(a) D_k has a binomial distribution (N_k, p_k). *In the model,* P_k is thus an estimator of p_k.

(b) p_k is itself random; we assume p_k to be independent and have the same distribution, with

$$E_\xi\, p_k = P,$$

$$\operatorname{Var}_\xi\, p_k = \sigma^2$$

for any k, in particular whatever the value of N_k.

In the model, after conditioning with p_k, we obviously have

$$E_\xi(D_k \mid p_k) = N_k\, p_k,$$

$$\operatorname{Var}_\xi(D_k \mid p_k) = N_k\, p_k(1 - p_k).$$

The *anticipated variance* of $\hat{P}$ is $E_\xi \operatorname{Var}\hat{P}$, to which we now turn our attention. For its evaluation, we denote

(a) $E_\xi(P_k - P)^2 = E_\xi(E_\xi(P_k - p_k + p_k - P)^2 \mid p_k)$

$$= \frac{P(1 - P) - \sigma^2}{N_k} + \sigma^2,$$

(b) $E_\xi P_k(1 - P_k) = E_\xi(E_\xi((P_k - P_k^2) \mid p_k))$

$$= E_\xi p_k(1 - p_k)\frac{N_k - 1}{N_k}$$

$$= (P(1 - P) - \sigma^2)\frac{N_k - 1}{N_k},$$

(c) $E_\xi Z_k Z_\ell = 0$, because of the independence of Z_k and Z_ℓ, clearing one extremely cumbersome term and $\pi_{k\ell}$.

When we combine all the pieces of (2.3) and (2.4), a minor algebraic miracle occurs, producing the expression

$$\boxed{\begin{gathered} E_\xi \text{Var}\, \hat{P} \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \\ \text{where } \tau^2 = P\,(1 - P) - \sigma^2 \\ \text{(by nature a positive quantity)} \end{gathered}} \;. \qquad (2.1.1)$$

Comment:

The algebraic miracle is easily explained if we are not seeking the variance in the sole context of sample design. It is in fact the result of a model slightly more general than the one suggested.

Suppose we wish to estimate the total $N\bar{Y} = \sum_U Y_i$ of a variable Y and suppose that, to this end, a two-stage sample is drawn: in the first stage, primary units k are drawn with π_k probability and, in the second, n_k final units are drawn by simple random sampling.

We are assuming a model in which:

$$Y_i = \bar{Y} + \alpha_k + \epsilon_i,$$

with α_k a variable linked to the PU of index k. α_k is independent, subject to the same zero expectation and has a variance σ^2. ϵ_i is also independent, centred and has a variance τ^2. With $\pi_i^* = \pi_k n_k / N_k$ (N_k = size of PU number k), the Horvitz-Thompson estimator of the *total* is $\hat{Y} = \sum Y_i / \pi_i^*$, the sum being extended to the sample. In the model, and conditionally in the sample, we have

$$\text{Var}_\xi(\hat{Y} \mid s) = \sum_s \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right).$$

For this expression, expectation is again expressed in the form of equation (2.1.1).

2.2 Search for an Optimum Sample Design

The maximum variance of $\hat{P}$ is set by the criteria selected for quality control. As the survey is repeated for each processing unit, it is only natural to seek to minimize the expected survey cost given in (2.1.1), *i.e.*

$$E \sum_s (C_o + n_k C_1) = \sum_U \pi_k (C_o + n_k C_1). \qquad (2.2.1)$$

The problem of optimization is expressed as:

$$\text{To minimize} \quad \sum_U \pi_k (C_o + n_k C_1)$$

with the constraints

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \leq V_o$$

and for any k, $n_k \leq N_k$.

Let us now apply a Lagrange multiplier λ to the first constraint - which will obviously be saturated - and multipliers μ_k to the others. We obtain the solutions

$$C_o + n_k C_1 = \lambda \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \qquad (2.2.2)$$

and, for any k:

$$C_1 \pi_k = \lambda \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \qquad (2.2.3)$$

with

$$\mu_k = 0 \quad \text{if} \quad n_k < N_k \quad \text{and} \quad \mu_k > 0 \quad \text{if} \quad n_k = N_k.$$

For the use of Lagrange multipliers, see for example Luenberger (1973).

For all primary units in which $\mu_k = 0$ (the largest), we obtain

$$n_k = \frac{\tau}{\sigma}\left(\frac{C_o}{C_1}\right)^{1/2} = n^*. \qquad (2.2.4)$$

Each primary unit receives the same allocation, which corresponds to the consistent accuracy principle. Going back to equation (2.2.3), we observe that, again for these primary units, the probability of inclusion π_k must be proportional to size N_k, *i.e.*

$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\tau}{n^*} N_k. \qquad (2.2.5)$$

This is the standard proof of a self-weighting one-stage survey in which the first stage is drawn with probabilities proportional to a measure of size. (See for example Cochran 1977).

Since n_k is independent of N_k, it is impossible to have $n_k = N_k$ or $\mu_k > 0$ unless $N_k \leq n^*$. Equation (2.2.2) gives us the probability of inclusion to within one factor:

$$\pi_k = \lambda^{1/2} N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k}\right)^{1/2} = \lambda^{1/2} N_k^{1/2} \left(\frac{N_k\sigma^2 + \tau^2}{N_k C_1 + C_o}\right)^{1/2}. \tag{2.2.6}$$

Relations (2.2.5), valid if $N_k \geq n^*$, and (2.2.6) valid if $N_k \leq n^*$, establish that π_k is proportional to a known variable $T_k = f(N_k)$, for which the graph is given in Figure 1.

To fully define the survey, the number m of primary units to be drawn must still be set. $T = \sum_U T_k$ is also a known quantity.

If we restrict ourselves to fixed size sampling, we have $\pi_k = m\, T_k/T$. m may be determined by importing this value into the variance constraint, *i.e.*

$$N^2 V_o m = T \sum_U \frac{N_k^2}{T_k}(\sigma^2 + \tau^2/n_k).$$

If, as a first approximation, assuming $T_k = N_k$, we obtain the simplified form:

$$m\, V_o = \sigma^2 + \tau^2/n^*.$$

We now have a full solution to the problem.

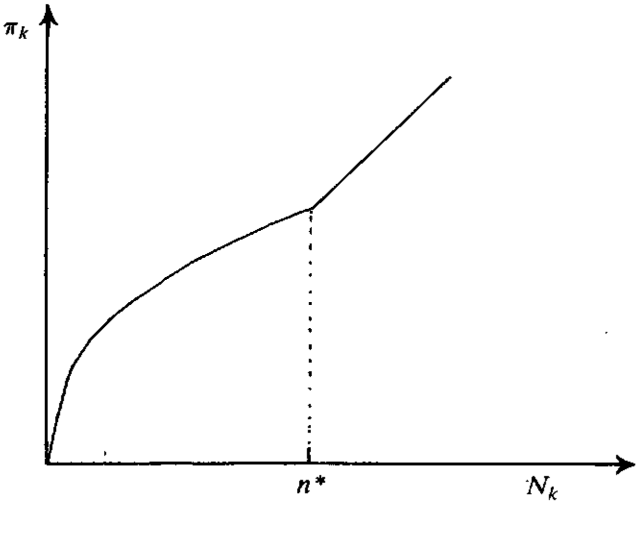

Figure 1. Graph of π_k as a function of N_k

3. OPTIMUM ESTIMATE FOR A TWO-STAGE SURVEY IN WHICH THE PRIMARY UNITS ARE STRATIFIED

The harsh facts of the situation complicate the problem somewhat: because a number of types of forms must be controlled separately, a fairly general problem, described below, arises.

For each primary unit (a district in a processing unit) we know the population N_{kg} of secondary units belonging to G groups. The "population" of PU number k is $N_{k+} = \sum_g N_{kg}$; that of group g is $N_{+g} = \sum_k N_{kg}$. As described above, we are looking for the probability of inclusion π_k with which to sample PU number k, the number of PUs to be drawn and the allocation n_{kg} of the sample among the various groups in PU k, knowing that these n_{kg} units are drawn by simple random sampling from among the N_{kg} units available.

3.1 Search for an Optimum Model Assisted Design

In each group, we postulate a model identical to the one formulated in section (2.1) (or the more general form described in the comment on that section).

For $g = 1$ to G, we have therefore:

$$v_g = E_\xi \operatorname{Var}(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k}(\sigma_g^2 + \tau_g^2/n_{kg}). \tag{3.1.1}$$

The cost function is expressed in the general form (1.2). We are seeking to minimize the expected survey cost

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g\right), \tag{3.1.2}$$

under constraints $V_g \leq \mathcal{V}_g$, where quantities $\mathcal{V}_g$ are externally fixed (*e.g.* quality of data to be obtained, tightness of control.

In this form, the problem can prove fairly complex. We write a general form of a Lagrange multiplier:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

The problem sets $\lambda = 1$, λ_g being multipliers to be determined. In a simple variant, values are set for λ_g: we wish to minimize a given linear combination of variances under a cost constraint. In all the hypotheses, by differentiation with respect to n_{kg} (considered a real variable), we obtain

$$\lambda \pi_k^2 C_g = \lambda_g N_{+g}^{-2} N_{kg}^2 \tau_g^2 / n_{kg}^2. \tag{3.1.3}$$

π_k being for the moment to be defined to within one factor, we may write

$$\pi_k\, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}}. \qquad (3.1.4)$$

By summing over k, we deduce that

$$E\, n_{+g} = \sum_U \pi_k\, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g. \qquad (3.1.5)$$

The total size of the sample in each group is thus directly linked to multiplier λ_g.

Differentiation of the Lagrange multiplier with respect to π_k gives us new relations which, when combined with (3.1.4), are miraculously simplified to give

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g}\right)^2 n_{kg}^2, \qquad (3.1.6)$$

or, if we introduce the numbers

$$n_g^* = \left(\frac{C_o}{C_g}\right)^{1/2} \frac{\tau_g}{\sigma_g},$$

we write

$$\sum_g \left(\frac{n_{kg}}{n_g^*}\right)^2 = 1. \qquad (3.1.7)$$

As may be seen in equation (2.2.4), n_g^* is the number of secondary units to be drawn per PU if there is a single group; n_{kg} is always less than n_g^*.

From (2.1.4), (3.1.5) and (3.1.7) we obtain the relations:

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g\, \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}}\right)^2. \qquad (3.1.8)$$

Thus, π_k is proportional to T_k such that $T_k^2 = \sum_g \lambda_g \sigma_g^2 N_{kg}^2 / N_{+g}^2$, which appears to be a satisfactory measure of size. The relations (3.1.4) show that, if k is fixed, n_{kg} is proportional to $n_g^*\, \lambda_g^{1/2}\, \sigma_g\, N_{kg}/N_{+g}$; taking into account (3.1.7), we obtain

$$n_{kg} = n_g^*\, \lambda_g^{1/2}\, \sigma_g \frac{N_{kg}}{N_{+g}}\, T_k^{-1}. \qquad (3.1.9)$$

3.2 Explicit Solutions to Two Specific Cases

(a) If λ_g were known, *i.e.* if $\sum_g \lambda_g v_g$ were minimized under a cost constraint, then (3.1.2) and (3.1.9) could be used to calculate T_k. By transfering

$$\pi_k = m\, T_k / T \left(T = \sum_U T_k,\ m \ \begin{matrix}\text{number of primary}\\ \text{units to be drawn}\end{matrix}\right)$$

to budget constraint $C_T \le C_T^*$, we find that

$$C_T^* = \frac{m}{T} \sum_U \left(C_o T_k + \sum_g C_g\, n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}}\right)$$

i.e.

$$m = C_T^* \Big/ \left(C_o + \sum_g C_g\, n_g^* \cdot \frac{\lambda_g^{1/2} \sigma_g}{T}\right).$$

If a single λ_g is not equal to zero, it is fairly easy to check that the result is the one given at the end of section (2.2).

(b) The initial problem (min C_T under $V_g \le \mathcal{V}_g$) is resolved fairly easily in two specific cases.

b1 - *Maximum dispersion* of the groups. For any PU k, we have $N_{kg} = N_{k+}$ for a given k. The problem is broken down into G separate problems, each being of the type examined in section 2.

b2 - *Minimum dispersion*. The distribution is the same in all the PUs; in other words, for any k and any g, we have

$$N_{kg} = N_{k+} \frac{N_{+g}}{N} \quad \text{with} \quad \left(N = \sum_g N_{+g}\right),$$

T_k is then proportional to N_{k+}, and n_{kg} is quantity $n_g^* u_g$ independent of k.

With $\pi_k = m N_{k+}/N$, we obtain by writing $V_g = \mathcal{V}_g$:

$$m \mathcal{V}_g = \sigma_g^2 + \tau_g^2 / n_g^* u_g$$

i.e.

$$m = \frac{\sigma_g^2}{\mathcal{V}_g} + u_g^{-1} \frac{\tau_g^2}{n_g^* \mathcal{V}_g}.$$

Thus we obtain G-1 linear relations between the u_g^{-1}, in principle permitting full resolution of the problem, knowing that the sum of u_g^2 is equal to 1.

3.3 A Numerical Algorithm for Determining the Optimum Solution to the General Case

An iterative numerical resolution of the problem may be achieved as follows.

Step 1: An approximate sample allocation is set in each group (n_{+g} units in group g). The process may be facilitated by using the approximate solution based on the hypotheses in point (a) or point (b).

Step 2: The value of λ_g is determined from relations (3.1.5):

$$\lambda_g = C_g\, n_{+g}^2 / \tau_g^2.$$

Step 3: π_k is determined from relations (3.1.8). Specifically, the sum of π_k sets the number of PUs to be drawn.

Step 4: n_{kg} is determined from relations (3.1.4). Subsequent iteration is possible by returning to step 2, in the expectation that the algorithm will converge toward the optimization solution.

Comment: The probability of drawing a type g unit is

$$\pi_k n_{kg} \Big/ N_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g / N_{+g}.$$

Because it does not depend on primary unit k, it is the same for each unit in a given group g (equal probability survey). Size n_{+g}, or at least its mathematical expectation, may be deduced from the sample in group g. In practice, sample size is sometimes set arbitrarily: this entails determining λ_g or, implicitly, variances $\mathcal{V}_g$. This is another fairly common result.

4. OPTIMUM ESTIMATE ASSISTED BY A MEASURE OF THE DIFFICULTY OF CODING A RECORD

The task is to estimate the proportion of forms containing a coding error in universe U of all forms coded in a given week by one regional branch. The problem is identified by the following characteristic: because all IFs are precoded, it is possible, using information drawn from the trial census, to attribute to each one a positive numerical variable X_i representing its "difficulty". This variable is calibrated in such a way that Y_i (equal to 1 if there is an error and 0 if there is not) has an "expectation" proportional to X_i.

The same cost control considerations suggest a two-stage survey.

- In the first stage of the survey, we draw a sample s_1 of districts k (primary units), with π_k unequal probabilities to be determined. $\pi_{k\ell}$ denotes the probability of inclusion, double in value in this instance.
- In the second stage of the survey, a sample s_k of final units (forms) in primary unit sample k is drawn. $\pi_{i|k}$ denotes the probability of inclusion of the unit in primary unit k, $\pi_{ij|k}$ the probability of inclusion of the pair (i,j) in the primary unit; and $s = U_{k \epsilon s_1} s_k$, the sample of final units.

$X_k = \sum_{i \epsilon k} X_i$ denotes the total of X_i in primary unit k, $X = \sum_{k \epsilon U_0} X_k = \sum_U X_i$ and similar notations are used for all the variables. (U_o denotes the population of primary units - districts, U the population of final units - forms).

The aim is to estimate a quantity in the form $R = \sum_U Y_i / \sum_U W_i$ where W_i is a known variable for each form. This may be $W_i = 1$ or $W_i = X_i$, whichever measure of the error rate seems the more satisfactory.

4.1 Selection of Estimator and Variance

(a) For primary unit k, the total Y_k of the Y_i for $i \epsilon k$ is commonly estimated by the ratio

$$\hat{Y}_k = X_k \left(\sum_{s_k} Y_i / \pi_{i|k} \right) \Big/ \left(\sum_{s_k} X_i / \pi_{i|k} \right) = X_k \hat{a}_k$$

where $\hat{a}_k$ estimates $a_k = Y_k / X_k$ with a slight biais.

(b) To estimate ratio Y/X, we use

$$\hat{a} = \frac{\sum_{s_1} \frac{\hat{Y}_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}.$$

(c) If we wish to estimate R, we note that

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

where X and W are known totals (*e.g.* total difficulty, total number of forms). As variable X_i was selected for its good correlation with Y_i, an *a priori* valuable estimator of R is

$$\hat{R} = \hat{a} \frac{X}{W}$$

and the only real question concerns the estimate of $a = \sum_k a_k X_k / X$.

(d) we have

$$\mathrm{Var}(\hat{a}) = \mathrm{Var}\, E(\hat{a} \mid s_1) + E\, \mathrm{Var}(\hat{a} \mid s_1).$$

For the first term, taking into account the fact that $\hat{a}_k$ is an approximate unbiased estimator of a_k, we may write

$$\mathrm{Var}\, E(\hat{a} \mid s_1) \simeq \frac{1}{X^2} \mathrm{Var}\left(\sum_{s_1} \frac{(a_k - a) X_k}{\pi_k} \right)$$

$$= \frac{1}{X^2} \left(\sum_k \frac{(a_k - a)^2 X_k^2}{\pi_k^2} \right.$$

$$\left. + \sum\sum_{k \neq l} (a_k - a)(a_\ell - a) \frac{X_k X_\ell \pi_{k\ell}}{\pi_k \pi_\ell} \right). \quad (4.1.1)$$

For the second term, *conditional on* s_1, we have

$$\mathrm{Var}\left(\frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}\right) = \left(\sum_{s_1} \frac{X_k}{\pi_k}\right)^{-2} \cdot \sum_{s_1} \mathrm{Var}(\hat{a}_k) \frac{X_k^2}{\pi_k^2}.$$

For this quantity, the expectation is approximately

$$X^{-2} \sum_k \mathrm{E}\ \mathrm{Var}(\hat{a}_k \mid s_1) \frac{X_k^2}{\pi_k}, \qquad (4.1.2)$$

with

$$\mathrm{Var}(\hat{a}_k \mid s_1) = \mathrm{Var} \frac{\sum_{s_k} \frac{Y_i}{\pi_{i|k}}}{\sum_{s_k} \frac{X_i}{\pi_{i|k}}} \simeq \frac{1}{X_k^2} \mathrm{Var} \sum_{s_k} \frac{Y_i - a_k X_i}{\pi_{i|k}}$$

$$= \frac{1}{X_k^2}\left(\sum_{i \epsilon k} \frac{(Y_i - a_k X_i)^2}{\pi_{i|k}}\right.$$

$$\left. + \sum_{k \neq l}\sum \frac{(Y_i - a_k X_i)(Y_j - a_k X_j)\pi_{ij|k}}{\pi_{i|k}\,\pi_{j|k}}\right).$$

As in the preceding sections, we arrive at formulae that are complex and, in the final analysis, unusable. A model will simplify things somewhat.

4.2 Introduction of a Model

The model has the same structure as those used previously:

(a) a_k is an independent random variable with the same expectation and the same variance:

$$E_\xi\, a_k = a \qquad \mathrm{Var}_\xi\, a_k = \sigma^2.$$

The variance takes into account operator influence, which we make no attempt to isolate, and also such factors as day of the week, time of day, day of the month *etc.*...

(b) Conditional on a_k, Y_i in primary unit k is an independent Bernoulli variable with $E_\xi(Y_i \mid k) = a_k X_i$

$$\mathrm{Var}_\xi(Y_i \mid k) = a_k X_i - a_k^2 X_i^2.$$

Comment:

Variable X_i, which has no actual concrete meaning, is defined to within one factor of scale. Conversely aX_i and σX_i, being probabilities, have an invariant physical interpretation. In what follows, one must always keep in mind that the results are invariant if X_i is multiplied by an arbitrary factor, on condition that a and σ are divided by the same factor. $\mathrm{Var}(\hat{a})$ in particular has no concrete meaning; $\mathrm{Var}(\hat{a}X)$ is an exception.

As before, we examine anticipated variance, expectation under the model of the sum of (4.1.1) and (4.1.2).

For the first term, the expectation of the cross products is of course zero. The expectation under the model for this term is thus:

$$X^{-2}\sigma^2 \sum_k \frac{X_k^2}{\pi_k}.$$

For the second term, we find (in light of the definitions given in 4.2.a and 4.2.b)

$$X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i E_\xi \frac{(a_k X_i - a_k^2 X_i^2)}{\pi_{i|k}}$$

$$= X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{aX_i - (a^2 + \sigma^2) X_i^2}{\pi_{i|k}}.$$

Therefore, overall

$$E_\xi\, \mathrm{Var}(\hat{a}X) = \sigma^2 \sum_{k \epsilon U_o} \frac{X_k^2}{\pi_k}$$

$$+ \sum_{k \epsilon U_o} \frac{1}{\pi_k} \sum_{i \epsilon k} \frac{aX_i - (a^2 + \sigma^2) X_i^2}{\pi_{i|k}}.$$

No algebraic miracle occurs here. *For simplification*, we assume that $(a^2 + \sigma^2)X_i^2$ is negligible in the face of aX_i. Numerically, we may expect $aX_i = 2 \text{ to } 5 \times 10^{-2}$ and $(a^2 + \sigma^2)X_i^2 = 3 \text{ to } 30 \times 10^{-4}$: whence the approximation

$$E_\xi\, \mathrm{Var}(\hat{a}X) \simeq \sigma^2 \sum_{k \epsilon U_o} \frac{X_k^2}{\pi_k} + a \sum_{k \epsilon U_o} \frac{1}{\pi_k} \sum_{i \epsilon k} \frac{X_i}{\pi_{i|k}}.$$

4.3 Sample Design Optimization

We use the following cost function:

$$C = \sum_{s_1} (C_o + C_1 n_k).$$

Here, $n_k = \sum_{i\epsilon k} \pi_{i|k}$ is the size of the sample drawn in district k (supposedly set at fixed size s_1). Its expectation is

$$C_T = \sum_{k\epsilon U_o} \pi_k (C_o + C_1 n_k).$$

Let

$$\pi_{i|k} = n_k P_i \left(\text{with } \sum_{i\epsilon k} P_i = 1\right) \quad \text{and} \quad Q_k = \pi_k n_k.$$

The problem of optimization is now

$$\text{Min: } C_o \sum_k \pi_k + C_1 \sum_k Q_k$$

$$\text{under: } \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i\epsilon k} \frac{X_i}{P_i} \leq \mathcal{V}_o.$$

In this form, we are pleased to observe that the terms in $\sum_i X_i/P_i$ may be minimized independently of the other terms. In other words, n_k has no impact on this term. Leaving optimization of the second stage of the survey until later, S_k^{*2} denotes the optimized value of $\sum_i X_i/P_i$.

With a Lagrange multiplier λ, by differentiation with respect to π_k and Q_k, we obtain

$$*C_o = \lambda \sigma^2 \frac{X_k^2}{\pi_k^2} \quad \text{i.e.} \quad \boxed{\pi_k \text{ proportional to } X_k} \quad (4.3.1)$$

$$*C_1 = \lambda a \frac{S_k^{*2}}{Q_k^2} \quad \text{whence} \quad \boxed{n_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X_k}}. \quad (4.3.2)$$

Specifically, the primary units are drawn with probabilities proportional to total difficulty, a standard resolution (see for example Särndal, Swensson, Wretman, 1992, Chapter 12).

We now move on to sub-district sampling (second stage of survey).

Beginning with a simple, straightforward case, forms are drawn one by one. Minimization produces P_i proportional to $\sqrt{X_i}$. A simple calculation shows that $S_k^* = \sum_{i\epsilon k} \sqrt{X_i}$. We can now calculate n_k using (4.3.2), and our problem is fully resolved.

In practice, things are more complicated. For fairly obvious reasons, only forms for entire households are selected. In other words, the second stage of the survey is a *cluster* survey. The values of P_i are the same (*i.e.* P_m) for all the members of a given cluster (household) m.

Let X_m be the sum of X_i individuals i in household m. The problem is to minimize $\sum X_m/P_m$ under $\sum n_m P_m = 1$, with n_m the size of household m. We easily reach solution

$$P_m = \sqrt{\bar{X}_m} \Big/ \sum n_m \sqrt{\bar{X}_m},$$

with $\bar{X}_m = X_m/n_m$, *mean difficulty of forms IF in household m*. From this we determine $S_k^* = \sum n_m \sqrt{\bar{X}_m}$.

This solution enables us to determine the number n_k of *final units* to be drawn using (4.3.2). However, the number of clusters (*households*) has not been determined: this snag was predictable. In fact, the cost function does not imply this constraint. To obtain the number m_k of clusters to be drawn, we arrange matters so that the expectation of the number of final units is equal to n_k. Thus,

$$m_k \left(\sum n_m \sqrt{\bar{X}_m} \right) \Big/ \sum \sqrt{\bar{X}_m}$$

whence

$$m_k = n_k \frac{\sum \sqrt{\bar{X}_m}}{\sum n_m \sqrt{\bar{X}_m}}.$$

Taking into account (4.3.2), we also have

$$m_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{\sum \sqrt{\bar{X}_m}}{X_k}$$

and the probability a given household being drawn is thus

$$\frac{m_k \sqrt{\bar{X}_m}}{\sum \sqrt{\bar{X}_m}}.$$

Following a number of algebraic manipulations, the value of the optimum variance is found to be:

$$E_\xi \operatorname{Var} (\hat{a}X)_{\text{OPT}} = \frac{(\sigma X)^2}{m} \left(1 + \frac{a}{\sigma} \frac{a^{-1/2} S^*}{X} \left(\frac{C_1}{C_o}\right)^{1/2}\right).$$

This form respects the homogeneous character of the different factors. In particular, we have $a^{-1/2} S^*/X = a^{1/2} S^*/aX$: the denominator may be interpreted as total number of errors in a lot; the numerator is homogeneous for a given size.

We now have a full solution to the problem.

Comment 1:

In both cases discussed, S_k^* is multiplied by $C^{1/2}$ if X_i is multiplied by C. The formula that gives n_k is thus invariant on the scale of measurement.

Comment 2:

The solution that entails drawing clusters favours small clusters made up of final units with a high index of difficulty.

Comment 3:

As in preceding sections, we determine the probability of single selection, but not the probability of dual selection. Therefore the algorithm for the draw, which sets the latter, has no influence. This is quite common, keeping in mind that the complementary data used to optimize the draw determines π_k and $\pi_{i|k}$ but have no influence on dual probabilities.

5. APPLICATIONS TO CONTROL BY SURVEY OF THE QUALITY OF THE 1990 FRENCH CENSUS

5.1 Problem of Data Capture Control

The sampling techniques described in sections 2 and 3 were designed to control data capture for the 1990 Census. A brief description of the operation would enhance understanding of the nature of the statistical problems involved.

The basic collection unit is the district, which corresponds, in a city, to a block of houses and, in the country, to a village or group of hamlets. It covers a population that ranges from zero inhabitants to approximately 2,000 (the mean values are 150 dwellings and approximately 350 inhabitants).

When collection is completed and the results are audited, the various census forms (specifically individual forms (IF) and dwelling forms (DF)) are meticulously counted for each district. The summary data for a district are computerized; the forms themselves, collated into district files, are forwarded to data capture.

Groups of districts comprising approximately 100,000 dwellings are constructed. The processing units (PU) are processed for INSEE by contractors. INSEE, the "client" in terms of control theory, monitors the quality of each contractor's work by sampling a specific number of forms in each PU.

The aim of the survey described in paragraph 2 is to estimate, to an accuracy (standard deviation) of one point, the proportion of forms containing an error in each PU. The maximum proportion of forms containing an error cannot exceed 4%. A trial census covering approximately 400 districts allows for an estimate of the values of the two model parameters. We find:

$$\sigma^2 \simeq P^2 \simeq 14.10^{-4}$$

$$\tau^2 \simeq P \simeq 4.10^{-2}.$$

Cost function (1.1) is assessed in terms of working time. Based on on-site control measures, 5 minutes is the estimate of the time required to process one district folder (from the time it is taken from the shelf to the time it is returned there) and 30 seconds the estimate of the time required to process one IF. With the numerical data, design optimization based on the hypotheses in section 1 allows for control of 40 districts per processing lot and 16 forms per district.

After discussing the solution with the team responsible for the census, it emerged that two types of documents (individual forms (IF) and dwelling forms (DF)) were to be controlled. The first approximation had taken no account of the latter, which are less likely to contain errors and take only about half as long to code as IFs. However, some districts (*e.g.* a commune with a thriving tourist industry) contain a large majority of secondary dwellings, and so produce many DFs but very few IFs. Because the situation required in-depth study, the theory given in section 3 was developed.

In the case of the census, the number of groups G is equal to 2 ($g = 1$ for the IFs and $g = 2$ for the DFs). The numerical data for the two groups are:

$$\cdot\ P_1 = 0{,}04 \qquad \sigma_1 = P_1 \qquad \tau_1^2 = P_1(1 - P_1) - \sigma_1^2 = P_1 - 2P_1^2,$$

$$\cdot\ P_2 = 0{,}01 \qquad \sigma_2 = P_2 \qquad \tau_2^2 = P_2 - P_2^2,$$

$$\cdot\ \mathcal{V}_1 = (0{,}0075)^2 \qquad \mathcal{V}_2 = (0{,}0150)^2.$$

For the cost function, we selected $C_o = 5$ minutes, $C_1 = 0.5$ minute and $C_2 = 0.25$ minute. Optimization of the problem according to the hypotheses in section 3.2.b entailed examining 73 districts per processing unit. In practical terms, it meant processing 15 individual forms (and related DFs) for each district. For the districts that produce fewer than 15 IFs, all IFs were processed. For districts with zero IFs, 4 DFs were processed (if this number was less than the number of DFs in the district).

Comment:

The method described in part 2 seems to have a fairly broad field of application. One example: it was used to sample the 1992 French survey on migration of foreign nationals. For population centres with under 20,000 inhabitants, the sample was drawn in two stages. The first stage of the survey covered the 90 departments in which this type of population centre occurs. The foreign population (based on the census) was divided into 8 nationality groups, for which equally accurate indicators had to be found.

5.2 Problems Related to Coding

The second step in data preparation is known as operation COLIBRI (Codification en Ligne des Bulletins du Recensement des Individus). The operators in the regional branches of INSEE receive forms classified by district and code them for the 25% survey.

In practice, each operator works at a monitor that displays the identifier of the next dwelling to be included in the 25% sample, for which all IFs must be coded.

Coding quality is also controlled by survey. The control unit is all the work done in one week in a regional branch. The entire operation takes a little over one year in the 22 regional branches, and entails more than 1,000 surveys. The household is the unit to be controlled (*i.e.* all the IFs in a household drawn for inclusion in the control sample). The objective is to estimate the proportion of forms containing an error. This is done by automatic detection of forms in which there is a no match situation. The number of errors is determined by reconciliation. The control theory is discussed in section 4 of this paper. The index of difficulty of the forms was developed from the data captured for a study based on the previous census and by test. The procedure and results related to these control measures are described in detail in G. Badeyan (1992).

The practical and numerical application of the theory rests on hypotheses concerning the orders of magnitude of the different parameters (which requires linking them to a simple physical interpretation). In the census preparation phase, without accurate prior measurement, we used the values $\sigma/a = 0.5$ and $C_1/C_o = 0.1$.

Pursuant to a number of hypotheses concerning the other parameters, and after discussing the matter with experts, it was decided that the control would cover 50 districts, with approximately 20 IFs controlled in each one (by region and by week). Since model parameters can be re-estimated at any stage in the process, the initial order of magnitude can obviously be adjusted as the survey proceeds.

Final Comment:

The problem produces somewhat surprising results that are worthy of consideration.

In the first instance, as we assumed it would be possible to separate each form, the forms were drawn with a probability proportional to individual difficulty. We assumed, to some extent, that the cost of using individual information was zero.

In the second instance, the actual control process, it was assumed that cost was infinite and the only information with negligible cost was the information related to an entire household. The solution shows that the probability of drawing an individual (IF) as a function of the mean difficulty of coding the forms for the entire household of which the individual is a member.

The same phenomenon occurs in the district draw. If it is possible to separate the IFs, they are drawn with probabilities proportional to total difficulty; within a district, the difficult IF has a greater probability of selection. Conversely, suppose we are unable to separate IFs within a district. This will be the case, for example, if the designation of IFs to be controlled cannot be implemented in real time because of inadequate processing facilities. Districts would then be selected in proportion to mean difficulty: within a district, it would be necessary to proceed by simple random sampling.

In the first instance, the survey gives precedence to large districts, from which difficult IFs tend to be drawn. In the second instance, precedence is given to small difficult districts, from which forms are selected with equal probability. *In both instances*, we are seeking to increase the probability of surveying difficult IFs. The difference resides simply in the possibility (*i.e.* the cost) of collecting information when we need it.

ACKNOWLEDGEMENTS

The author would like to thank the Editor, the Associate Editor and the Referee for their extremely positive comments. He would also like to thank Claude Thelot, some of whose comments have been incorporated into this paper and Gérard Badeyan, who introduced the methods discussed here at INSEE. Last, but not least, he would like to thank Françoise Hitier, without whom this paper would never have seen the light of day.

REFERENCES

BADEYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, June 17 and 18, 1992, INSEE, Paris.

COCHRAN, W. (1977). *Sampling Techniques*, (3rd Edition). New York: Wiley.

DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.

LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 183–192
Statistics Canada

Conditional Properties of Post-Stratified Estimators Under Normal Theory

ROBERT J. CASADY and RICHARD VALLIANT[1]

ABSTRACT

Post-stratification is a common technique for improving precision of estimators by using data items not available at the design stage of a survey. In large, complex samples, the vector of Horvitz-Thompson estimators of survey target variables and of post-stratum population sizes will, under appropriate conditions, be approximately multivariate normal. This large sample normality leads to a new post-stratified regression estimator, which is analogous to the linear regression estimator in simple random sampling. We derive the large sample design bias and mean squared errors of this new estimator, the standard post-stratified estimator, the Horvitz-Thompson estimator, and a ratio estimator. We use both real and artificial populations to study empirically the conditional and unconditional properties of the estimators in multistage sampling.

KEY WORDS: Asymptotic normality; Regression estimator; Defective frames; Ratio estimator; Horvitz-Thompson estimator.

1. INTRODUCTION

1.1 Background

A major thrust in sampling theory in the last twenty years has been to devise ways of restricting the set of samples used for inference. In a purely design-based approach, as described in Hansen, Madow, and Tepping (1983), no such restrictions are imposed. Statistical properties are calculated by averaging over the set of all samples that might have been selected using a particular design. Although it is generally conceded that some type of design-based, conditional inference is desirable (Fuller 1981, Rao 1985, Hidiroglou and Särndal 1989), satisfactory theory has yet to be developed except in relatively simple cases. Alternative approaches are prediction theory, developed by Royall (1971) and many others, and the Bayesian approach, found in Ericson (1969), which avoid averaging over repeated samples through the use of superpopulation models. A design-based approach to conditioning was introduced by Robinson (1987) for the particular case of ratio estimates in sample surveys. Robinson applied large sample theory and approximate normality of certain statistics to produce a conditional, design-based theory for the ratio estimator.

In this paper, we extend that line of reasoning to the problem of post-stratification. Convincing arguments have been made in the past by Durbin (1969), Holt and Smith (1979) and Yates (1960) that post-stratified samples should be analyzed conditional on the sample distribution of units among the post-strata. However, as Rao (1985) has noted, the difficulties in developing an exact, design-based, finite sample theory for post-stratification in general sample designs may be intractable. Model-based, conditional analyses of post-stratified samples are presented in Little (1991) and Valliant (1993). The alternative pursued here is design-based and uses large sample, approximate normality in a way similar to that of Robinson (1987) as a means studying conditional properties of estimators.

1.2 Basic Definitions and Notation

The **target population** is a well defined collection of elementary (or analytic) units. For many applications the elementary units are either persons or establishments. We assume the target population has been partitioned into **first stage sampling units** (FSUs). For person based surveys the FSUs are commonly households, groups of households or even counties, while for establishment based surveys it is not uncommon that the individual establishment is an FSU. In any event, the collection of FSUs will be referred to as the **first stage sampling frame** (or just **sampling frame**). It is assumed that there are M FSUs in the sampling frame and they are labeled $1, 2, \ldots, M$. We also assume that the population units can be partitioned into K "post-strata" which can be used for the purposes of estimation.

We let y represent the value of the characteristic of interest (*e.g.* weekly income, number of hours worked last week, restricted activity days in last two weeks, *etc.*) for an elementary unit. Associated with the i^{th} FSU are $2K$ real numbers:

y_{ik} = aggregate of the y values for the elementary units in the i^{th} FSU which are in the k^{th} post-stratum,

N_{ik} = number of elementary units in the i^{th} FSU which are in the k^{th} post-stratum.

[1] Robert J. Casady and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001.

For each post-stratum we then define

$Y_{\cdot k} = \sum_{i=1}^{M} y_{ik}$ = aggregate of the y values for all elementary units in the k^{th} post-stratum,

$N_{\cdot k} = \sum_{i=1}^{M} N_{ik}$ = total number of elementary units in the k^{th} post-stratum.

In what follows we assume that the $N_{\cdot k}$ are known fixed values. In some surveys, the $N_{\cdot k}$ may actually be estimates themselves but our analysis is conditional on the set of $N_{\cdot k}$ used in estimation. In the Current Population Survey in the United States, for example, each $N_{\cdot k}$ is a population count projected from the previous decennial census using demographic methods. The population aggregate of the y values is given by $Y_{\cdot\cdot} = \sum_{k=1}^{K} Y_{\cdot k}$ and the total population size by $N_{\cdot\cdot} = \sum_{k=1}^{K} N_{\cdot k}$. In sections 1-3, we assume that the sampling frame provides "coverage" of the entire target population. In section 4, we consider the problem of a defective frame, *i.e.* one in which the coverage of the frame differs from that of the target population.

1.3 Sample Design and Basic Estimation

Suppose that the first stage sampling frame is partitioned into L strata and that a multi-stage, stratified design is used with a total sample of m FSUs. In the following, the subscript representing design strata is suppressed in order to simplify the notation. For the subsequent theory, it is unnecessary to explicitly define sampling and estimation procedures for second and higher levels of the design. However, for every sample FSU, we require estimators $\hat{y}_{ik}$ and $\hat{N}_{ik}$ so that $E_{2+}[\hat{y}_{ik}] = y_{ik}$ and $E_{2+}[\hat{N}_{ik}] = N_{ik}$ where the notation E_{2+} indicates the design-expectation over stages 2 and higher. Letting π_i be the probability that the i^{th} FSU is included in the sample and $w_i = 1/\pi_i$, it follows that the estimator $\hat{Y}_{\cdot k} = \sum_{i=1}^{m} w_i \hat{y}_{ik}$ is unbiased for $Y_{\cdot k}$ and the estimator $\hat{N}_{\cdot k} = \sum_{i=1}^{m} w_i \hat{N}_{ik}$ is unbiased for $N_{\cdot k}$.

1.4 An Analogue to Robinson's Asymptotic Result

Robinson (1987) studied the ratio estimator $(\bar{X}/\bar{x}_s)\bar{y}_s$ under simple random sampling with $\bar{y}_s$ being the sample mean of a target variable y, $\bar{x}_s$ being the sample mean of an auxiliary variable x, and $\bar{X}$ the population mean of x. Under certain conditions $(\bar{y}_s, \bar{x}_s)$ will be asymptotically, bivariate normal in large simple random samples. From Robinson's results it follows that the linear regression estimator $\bar{y}_s + \beta(\bar{X} - \bar{x}_s)$ is asymptotically design-unbiased conditional on $\bar{x}_s$. Results in this section extend that result to complex samples.

Following Krewski and Rao (1981), we can establish our asymptotic results as $L \to \infty$ within the framework of a sequence of finite populations $\{\Pi_L\}$ with L strata in Π_L. It should be understood that we implicitly assume (without formal statement) the sample design and regularity conditions as specified in Krewski and Rao and more fully developed in Rao and Wu (1985). Details of proofs add little to those in the literature and are omitted.

Converting to matrix notation, we let $Y = [Y_{\cdot 1} \ldots Y_{\cdot k}]'$, $N = [N_{\cdot 1} \ldots N_{\cdot k}]'$, $\hat{Y} = [\hat{Y}_{\cdot 1} \ldots \hat{Y}_{\cdot k}]'$, $\hat{N} = [\hat{N}_{\cdot 1} \ldots \hat{N}_{\cdot k}]'$ and $V = \text{var}\{[\hat{\hat{Y}}\ \hat{\hat{N}}]'\}$ where $\hat{\hat{Y}} = (1/N_{\cdot\cdot})\hat{Y}$ and $\hat{\hat{N}} = (1/N_{\cdot\cdot})\hat{N}$. Note that $\hat{\hat{Y}}$, which uses $N_{\cdot\cdot}$ in the denominator, is a notational convenience and does not estimate means in the post-strata. Analogous to conditions C4 and C5 of Krewski and Rao (1981), we assume that

$$\lim_{L \to \infty} \frac{Y_{\cdot k}}{N_{\cdot k}} = \mu_k, \quad \text{for} \quad k = 1, 2, \ldots, K, \tag{1}$$

$$\lim_{L \to \infty} \frac{N_{\cdot k}}{N_{\cdot\cdot}} = \phi_k > 0 \quad \text{for} \quad k = 1, 2, \ldots, K, \text{ and} \tag{2}$$

$$\lim_{L \to \infty} mV = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ (positive definite),} \tag{3}$$

where Σ is partitioned in the obvious manner. Note that we have again suppressed the subscript representing design strata. Assumptions (1)-(3) simply require that certain key quantities stabilize in large populations. Condition (2), in particular, assures that no post-stratum is empty as the population size increases. We now state the following.

Result: Assume the sample design and regularity conditions specified in Krewski and Rao and that Σ_{22}^{-1} exists; then, given $\hat{\hat{N}}$, the conditional distribution of $\hat{\hat{Y}}$ is asymptotically $\mathfrak{N}(M_1 + \Sigma_{12}\Sigma_{22}^{-1}(\hat{\hat{N}} - M_2), m^{-1}V_c)$, where $V_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $M_1 = \lim_{L \to \infty} \hat{\hat{Y}} = [\phi_1\mu_1 \ \ldots \ \phi_K\mu_K]'$ and $M_2 = \lim_{L \to \infty} \hat{\hat{N}} = [\phi_1 \ldots \phi_K]'$.

Proof. This result is analogous to the result for $K = 1$ given by Robinson (1987) and follows directly from the fact that the random vector

$$m^{1/2}\begin{bmatrix} \hat{\hat{Y}} - M_1 - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\hat{N}} - M_2) \\ \hat{\hat{N}} - M_2 \end{bmatrix}$$

tends in distribution to

$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_c & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right).$$

Strictly, as in Robinson, we consider the conditional distribution of $\hat{\hat{Y}}$ for $\hat{\hat{N}}$ in a cell of size $\epsilon m^{-1/2}$ for small ϵ. Note that in some sample designs $\mathbf{1}'\hat{N} = N_{\cdot\cdot}$ (such as those in which a fixed number of elementary units are selected with equal probabilities) in which case Σ_{22}^{-1} does not exist; in such cases only the first $K - 1$ post-strata are considered for the purpose of conditioning.

In the next section, the asymptotic mean of $\hat{\hat{Y}}$ is used to motivate a linear regression estimator of the population mean of the y's.

2. CONDITIONAL PROPERTIES OF ESTIMATORS FOR THE POPULATION MEAN

2.1 Estimators for the Population Mean

The **population mean** is, by definition,

$$\mu = \lim_{L \to \infty} (Y_{..}/N_{..}) = \lim_{L \to \infty} (\mathbf{1}' Y / \mathbf{1}' N) = \sum_{k=1}^{K} \phi_k \mu_k$$

where $\mathbf{1}'$ is a row vector of K ones. Note that the mean μ is not a finite population parameter but rather a limiting value. In large populations $(L \to \infty)$ μ and the actual finite population mean will be arbitrarily close. Four estimators of the population mean will be considered. The first three are standard estimators found in the literature while the fourth is a new estimator motivated by the asymptotic, joint normality of $\hat{\hat{Y}}$ and $\hat{\hat{N}}$:

(1) Horvitz-Thompson estimator

$$\hat{\hat{Y}}_{HT} = \mathbf{1}' \hat{Y} / \mathbf{1}' N = \mathbf{1}' \hat{\hat{Y}}.$$

(2) Ratio estimator

$$\hat{\hat{Y}}_{R} = \mathbf{1}' \hat{Y} / \mathbf{1}' \hat{N} = \mathbf{1}' \hat{\hat{Y}} / \mathbf{1}' \hat{\hat{N}}.$$

(3) Post-stratified estimator

$$\hat{\hat{Y}}_{PS} = N_{..}^{-1} \sum_{k=1}^{K} \left(\frac{N_{\cdot k}}{\hat{N}_{\cdot k}} \right) \hat{Y}_{\cdot k} = \boldsymbol{r}' \hat{\hat{Y}}$$

where

$$\boldsymbol{r}' = [N_{\cdot 1}/\hat{N}_{\cdot 1}, \ldots, N_{\cdot K}/\hat{N}_{\cdot K}].$$

(4) Linear regression estimator

$$\hat{\hat{Y}}_{LR} = [\mathbf{1}' (\hat{\hat{Y}} - \Sigma_{12} \Sigma_{22}^{-1} (\hat{\hat{N}} - M_2))].$$

The linear regression estimator is motivated by the form of the large sample mean of the conditional random variable $\hat{\hat{Y}} \mid \hat{\hat{N}}$ listed at the end of section 1.4 and is very similar to the generalized regression estimator discussed by Särndal, Swensson and Wretman (1992). The linear regression estimator (4) was also discussed in the context of calibration estimation by Rao (1992). It should be noted that the ratio estimator does not require that $N_{\cdot k}$ or their sum $N_{..}$ be known. The Horvitz-Thompson estimator only requires that $N_{..}$ be known, whereas the post-stratified and linear regression estimators require that $\{N_{\cdot k} \mid k = 1, \ldots, K\}$ be known. In practice, the linear regression estimator has the additional complication that the covariance matrices Σ_{12} and Σ_{22} are unknown and must be estimated from the sample. In implementing $\hat{\hat{Y}}_{LR}$ in section 3, the known finite population quantities $(1/N_{..})N$ will be used in place of the limiting vector M_2.

2.2 Conditional Expectations and Variances of the Estimators

Using the asymptotic setup given earlier, the expectations and variances of the four estimators can be computed conditional on $\hat{\hat{N}}$. For the case of post-stratification, conditioning on $\hat{\hat{N}}$ in a complex design is a natural extension of conditioning on the achieved post-stratum sample sizes in a simple random sample. In other situations, however, the question of what to condition on is a difficult one that may not have a unique answer (*e.g.*, see Kiefer 1977). First, define the following three matrices:

$$\boldsymbol{H} = \Sigma_{12} \Sigma_{22}^{-1},$$

$$\boldsymbol{R} = \boldsymbol{H} - \boldsymbol{D}(\mu), \quad \text{and}$$

$$\boldsymbol{P} = \boldsymbol{H} - \boldsymbol{D}(\mu_k),$$

where $\boldsymbol{D}(\mu) = \text{diag}(\mu, \ldots, \mu)$ and $\boldsymbol{D}(\mu_k) = \text{diag}(\mu_1, \ldots, \mu_k)$ are $K \times K$ diagonal matrices. Below, we state the mean and variance of the four estimators without providing any details of the calculations. When the sample of first-stage units is large, each of the estimators has essentially the same conditional variance. The Horvitz-Thompson, ratio, and post-stratified estimators are, however, conditionally biased, whereas the linear regression estimator is not. Thus, the linear regression estimator has the smallest asymptotic mean square error among the four estimators considered here. Rao (1992) also noted the optimality of the regression estimator within a certain class of difference estimators and its negligible large sample bias.

(1) Horvitz-Thompson estimator:

$$\text{E}[\hat{\hat{Y}}_{HT} \mid \hat{\hat{N}}] = \mu + [\mathbf{1}' \boldsymbol{H} (\hat{\hat{N}} - M_2)]$$

$$\text{var}[\hat{\hat{Y}}_{HT} \mid \hat{\hat{N}}] = m^{-1} [\mathbf{1}' (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}) \mathbf{1}]$$

$$= m^{-1} [\mathbf{1}' V_c \mathbf{1}] = V_{HT(c)}.$$

(2) Ratio estimator:

$$\text{E}[\hat{\hat{Y}}_{R} \mid \hat{\hat{N}}] = \mu + \left(\frac{N_{..}}{\hat{N}_{..}} \right) [\mathbf{1}' \boldsymbol{R} (\hat{\hat{N}} - M_2)]$$

$$= \mu + [\mathbf{1}' \boldsymbol{R} (\hat{\hat{N}} - M_2)] + o(m^{-1})$$

$$\text{var}[\hat{\hat{Y}}_{R} \mid \hat{\hat{N}}] = (N_{..}/\hat{N}_{..})^2 V_{HT(c)}$$

$$= V_{HT(c)} + o(m^{-(3/2)}).$$

(3) Post-stratified estimator:

$$\begin{aligned} \mathrm{E}[\hat{\bar{Y}}_{PS} \mid \hat{\bar{N}}] &= \mu + [\boldsymbol{r}'\boldsymbol{P}(\hat{\bar{N}} - \boldsymbol{M}_2)] \\ &= \mu + [\boldsymbol{1}'\boldsymbol{P}(\hat{\bar{N}} - \boldsymbol{M}_2)] + o(m^{-1}) \\ \mathrm{var}[\hat{\bar{Y}}_{PS} \mid \hat{\bar{N}}] &= m^{-1}[\boldsymbol{r}'\boldsymbol{V}_c\boldsymbol{r}] \\ &= V_{HT(c)} + o(m^{-(3/2)}). \end{aligned}$$

(4) Linear regression estimator:

$$\begin{aligned} \mathrm{E}[\hat{\bar{Y}}_{LR} \mid \hat{\bar{N}}] &= \mu \\ \mathrm{var}[\hat{\bar{Y}}_{LR} \mid \hat{\bar{N}}] &= V_{HT(c)}. \end{aligned}$$

As noted in section 1, some minor modifications of the above formulas are necessary for designs, such as simple random sampling, in which $\boldsymbol{1}'\hat{N}_{..} = N_{..}$. The derivation of the requisite modifications is straightforward and is not detailed here.

The large-sample biases of the first three estimators depend on $\hat{\bar{N}} - \boldsymbol{M}_2$. In other words, their biases are determined by how well the sample estimates the population distribution among the post-strata. In some special cases each of the first three can be conditionally unbiased. The post-stratified estimator, for example, will be approximately unbiased if $\boldsymbol{1}'(\boldsymbol{H} - \boldsymbol{D}(\mu_k)) = \boldsymbol{0}'$. This occurs in simple random sampling and is possible, though certainly not generally true, in more complex designs. The matrix $\boldsymbol{H}$ can be interpreted as the slope in a multivariate regression of $\hat{\bar{Y}}$ on $\hat{\bar{N}}$ or of $\bar{Y}$ on $\bar{N}$ when the sample estimates are close to the population values. Thinking heuristically in superpopulation terms, if $\mathrm{E}_\xi(y_{ik}) = \mu_k N_{ik}$, as in Valliant (1993), with E_ξ denoting an expectation with respect to the model, then $\mathrm{E}_\xi(Y_{.k}) = \mu_k N_{.k}$. The slope of the regression of $Y_{.k}$ on $N_{.k}$ is then μ_k and, in the unusual case in which the $\hat{\bar{Y}}_{.k}$'s are independent, $\boldsymbol{H}$ is diagonal. In fact $\boldsymbol{H} = \boldsymbol{D}(\mu_k)$, so the conditional design-bias of the post-stratified estimator would be zero. If, on the other hand, the model has an intercept, *i.e.* if $\mathrm{E}_\xi(Y_{.k}) = \alpha_k + \mu_k N_{.k}$, then the post-stratified estimator may have a substantial conditional design-bias. We will use this line of reasoning in the empirical study in section 3 to devise a population for which $\hat{\bar{Y}}_{ps}$ is conditionally biased.

Similar model-based thinking can be applied to the Horvitz-Thompson and ratio estimators to identify populations where the conditional design-biases will be predictably small for large samples. Suppose, as above, that the $\hat{\bar{Y}}_{.k}$'s are independent. If each post-stratum total is unrelated to the number of units in the post-stratum, *i.e.* a peculiar situation in which $\mathrm{E}_\xi(Y_{.k})$ does not depend on $N_{.k}$, then $\hat{\bar{Y}}_{HT}$ is conditionally design-unbiased. If $\mathrm{E}_\xi(Y_{.k}) = \mu N_{.k}$, implying that all elementary population units have the same mean regardless of post-stratum, then $\hat{\bar{Y}}_R$ is conditionally design-unbiased.

2.3 Unconditional Expectations and Variances of the Estimators

Unconditionally, all estimators are approximately design-unbiased as noted below. The relative sizes of the variances depend on the values of Σ_{12}, Σ_{22}, μ, and $\boldsymbol{D}(\mu_k)$. This is similar to the case of simple random sampling of a target y and an auxiliary x. In that case, whether the ratio estimator, $\bar{y}_s\bar{X}/\bar{x}_s$, or the regression estimator, $\bar{y}_s + b(\bar{X} - \bar{x}_s)$, has smaller design-variance also depends on the values of certain population parameters.

(1) Horvitz-Thompson estimator:

$$\begin{aligned} \mathrm{E}[\hat{\bar{Y}}_{HT}] &= \mu \\ \mathrm{var}[\hat{\bar{Y}}_{HT}] &= m^{-1}[\boldsymbol{1}'\Sigma_{11}\boldsymbol{1}]. \end{aligned}$$

(2) Ratio estimator:

$$\begin{aligned} \mathrm{E}[\hat{\bar{Y}}_{\mathrm{R}}] &= \mu + o(\mathrm{m}^{-1}) \\ \mathrm{var}[\hat{\bar{Y}}_{\mathrm{R}}] &= \mathrm{m}^{-1}[\boldsymbol{1}'[\Sigma_{11} - 2\mu\Sigma_{21} + \mu^2\Sigma_{22}]\boldsymbol{1}] \\ &\quad + o(\mathrm{m}^{-(3/2)}). \end{aligned}$$

(3) Post-stratified estimator:

$$\begin{aligned} \mathrm{E}[\hat{\bar{Y}}_{\mathrm{PS}}] &= \mu + o(\mathrm{m}^{-1}) \\ \mathrm{var}[\hat{\bar{Y}}_{\mathrm{PS}}] &= \mathrm{m}^{-1}[\boldsymbol{1}'[\Sigma_{11} - 2\boldsymbol{D}(\mu_{\mathrm{k}})\Sigma_{21} \\ &\quad + \boldsymbol{D}(\mu_{\mathrm{k}})\Sigma_{22}\boldsymbol{D}(\mu_{\mathrm{k}})]\boldsymbol{1}] + o(\mathrm{m}^{-(3/2)}). \end{aligned}$$

(4) Linear regression estimator:

The unconditional expectation and variance are the same as the conditional expectation and variance.

3. SIMULATION RESULTS

The theory developed in the preceding sections was tested in a set of simulation studies using three separate populations. The population size and basic sample design parameters for the three studies are listed in Table 1. The first population consists of a subset of the persons included in the first quarter sample of the 1985 National Health Interview Survey (NHIS) and the second population consists of a subset of the persons included in the September 1988 sample from the Current Population Survey (CPS). Both the NHIS and CPS are sample surveys conducted by the U.S. government. The variable of interest for the NHIS population is the number of restricted activity days in the two weeks prior to the interview and the variable of interest for the CPS population is weekly wages per person.

Table 1
Population Size and Basic Sample Design Parameters for Three Simulation Studies

Population	Pop. Size N	No. of FSUs M	No. of sample FSUs m
HIS	2,934	1,100	115
CPS	10,841	2,826	200
Artificial	22,001	2,000	200

Post-strata in the NHIS and CPS populations were formed on the basis of demographic characteristics (as is typically done in household surveys) in order to create population sub-groups that were homogenous with respect to the variable of interest. For the NHIS population the variables age and sex were used to define 4 post-strata and for the CPS population the variables age, race, and sex were used to define 8 post-strata.

The third population is artificial; it was created with the intention of producing a substantial conditional bias in the post-stratified estimator of the mean. As noted in section 2.2, $\hat{Y}_{PS}$ will be conditionally biased if the FSU post-stratum totals for the variable of interest, conditional on the number of units in each FSU/post-stratum, follow a model with a non zero intercept. With this in mind, we generated the population in such a way that

$$\mathrm{E}(y_{ik} \mid N_{ik}) = \alpha_k + \beta N_{ik} + \gamma N_{ik}^2, \tag{4}$$

where N_{ik} is the number of units in the k^{th} post-stratum for the i^{th} FSU and α_k, β and γ are constants. Specifically, five post-strata were used with $\alpha_k = 100k$ $(k = 1, \ldots, 5)$, $\beta = 10$ and $\gamma = -.05$. In total two thousand FSUs were generated with the total number of units in the i^{th} FSU, say $N_{i\cdot}$, being a Poisson random variable with mean 10. Then, conditional on $N_{i\cdot}$, the numbers of units in the five post-strata (*i.e.*, $N_{i1}, N_{i2}, \ldots, N_{i5}$) for the i^{th} FSU were determined using a multinomial distribution with parameters $N_{i\cdot}$ and $p_k = .20$ for $k = 1, 2, \ldots, 5$.

For FSUs having $N_{ik} \geq 1$, the value of the variable of interest for the j^{th} unit in the k^{th} post-stratum for the i^{th} FSU was a realization of the random variable

$$y_{ijk} = \alpha_k / N_{ik} + \beta + \gamma N_{ik} + \epsilon_{1i} + \epsilon_{2ik} + \epsilon_{3ijk} N_{i\cdot}$$

$$(j = 1, \ldots, N_{ik}; N_{ik} \geq 1),$$

where ϵ_{1i}, ϵ_{2ik} and ϵ_{3ijk} are three independent standardized chi-square (6 d.f.) random variables. This structure implies that $\mathrm{E}(y_{ik} \mid N_{ik})$ is given by (4). Furthermore, the values of the variable of interest for units within an FSU are correlated and the correlation depends upon whether the units are in the same post-stratum or not. This same algorithm was used in each of the 100 design strata. Twenty FSUs were generated in each design stratum giving a total of 2,000 FSUs.

A single-stage stratified design was used for the NHIS population with "households" being the FSUs. Ten design strata were used and an approximate 10% simple random sample of households was selected without replacement from each stratum. Each sample consisted of 115 households and each sample household was enumerated completely. A total of 5,000 such samples was selected for the simulation study.

Two-stage stratified sample designs were used for both the CPS and artificial populations. For the CPS population, geographic segments, employed in the original survey and composed of about four neighboring households, were used as FSUs and persons were the second-stage units. In both populations, 100 design strata were created with each stratum having approximately the same number of FSUs and a sample of $m = 2$ FSUs was selected with probability proportional to size from each stratum using the systematic sampling method described by Hansen, Hurwitz and Madow (1953, p. 343). Thus, 200 FSUs were selected for both populations. Second stage selection was also similar for both populations. For the CPS population a simple random sample of 4 persons was selected without replacement in each sample FSU having $N_{i\cdot} > 4$ and all persons were selected in each sample FSU where $N_{i\cdot} \leq 4$. For the artificial population the within FSU sample size was set at 15 rather than 4 which resulted in the complete enumeration of most sample FSUs. A total of 5,000 samples were selected from each of the populations for the simulation study.

In each sample, we computed $\hat{Y}_{HT}$, $\hat{Y}_R$, $\hat{Y}_{PS}$ and two versions of $\hat{Y}_{LR}$. For the first version of the regression estimator, denoted $\hat{Y}_{LR}$(emp) in the tables, $\boldsymbol{H}$ was estimated separately from each sample as would be required in practice. Each component of Σ_{12} and Σ_{22} was estimated using the ultimate cluster estimator of covariance, appropriate to the design, as defined in Hansen, *et al.* (1953, p.419). The second version, denoted $\hat{Y}_{LR}$(theo), used the same value of $\boldsymbol{H}$ in each sample, which was an estimate more nearly equal to the theoretical value of the $\boldsymbol{H}$ matrix. For the CPS and artificial populations, the theoretical $\boldsymbol{H}$ matrix was estimated from empirical covariances derived from separate simulation runs of 5,000 samples. For the NHIS population the design was simple enough that a direct theoretical calculation of $\boldsymbol{H}$ was done. As the sample of FSUs becomes large, the performance of $\hat{Y}_{LR}$(emp) should approach that of $\hat{Y}_{LR}$(theo). The performance of $\hat{Y}_{LR}$(theo) is, consequently, a gauge of the best that can be expected from the empirical version of the regression estimator for a given sample size.

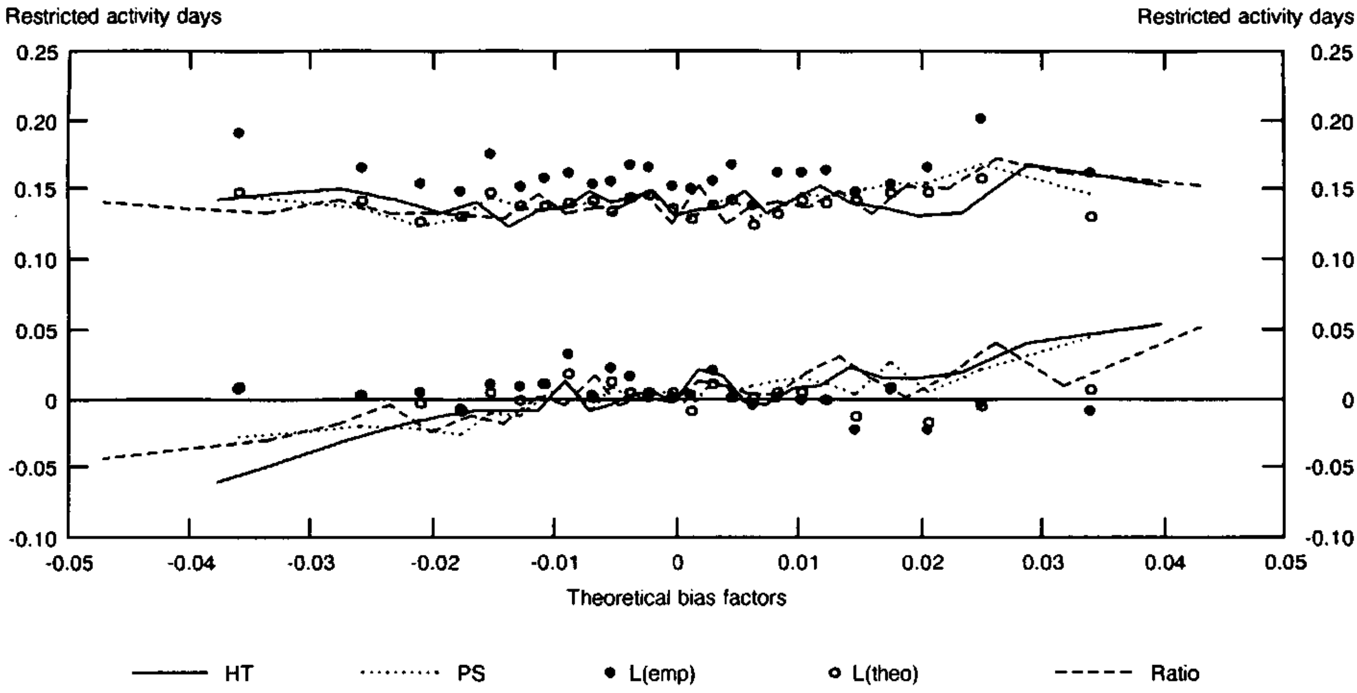

Figure 1. HIS simulation, $m = 115$

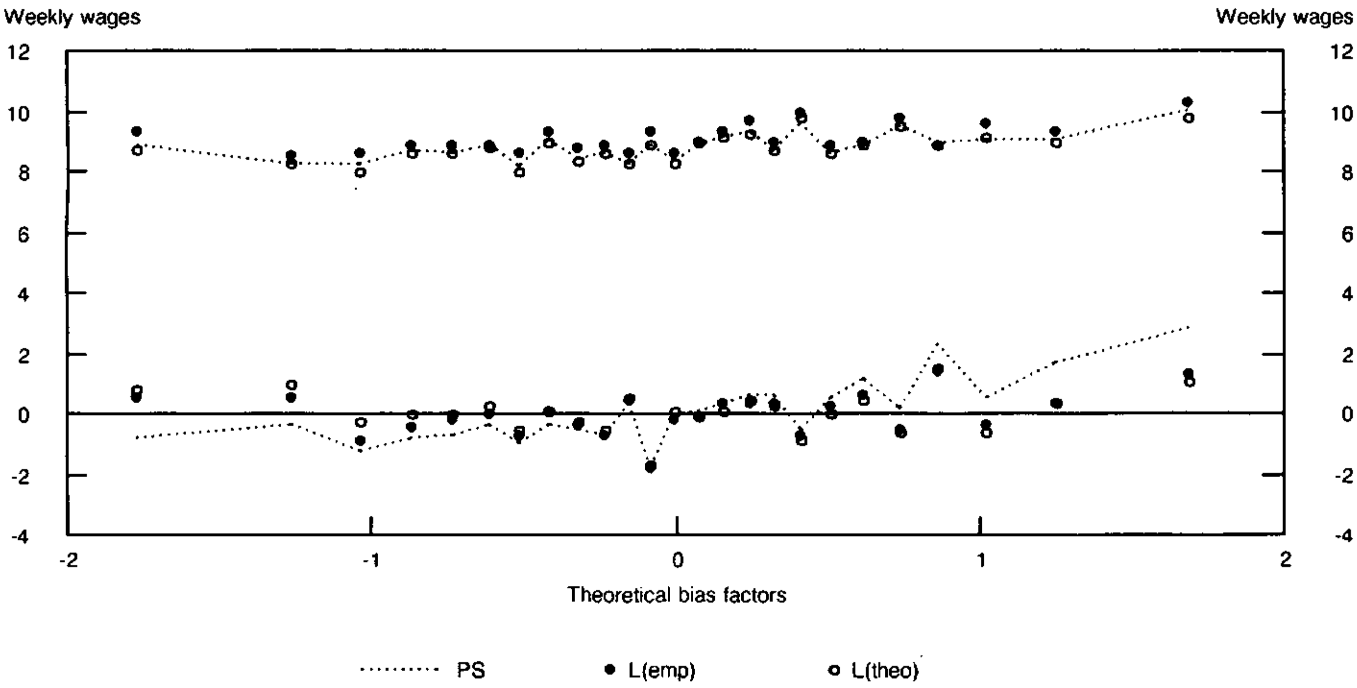

Figure 2. CPS simulation, $m = 200$

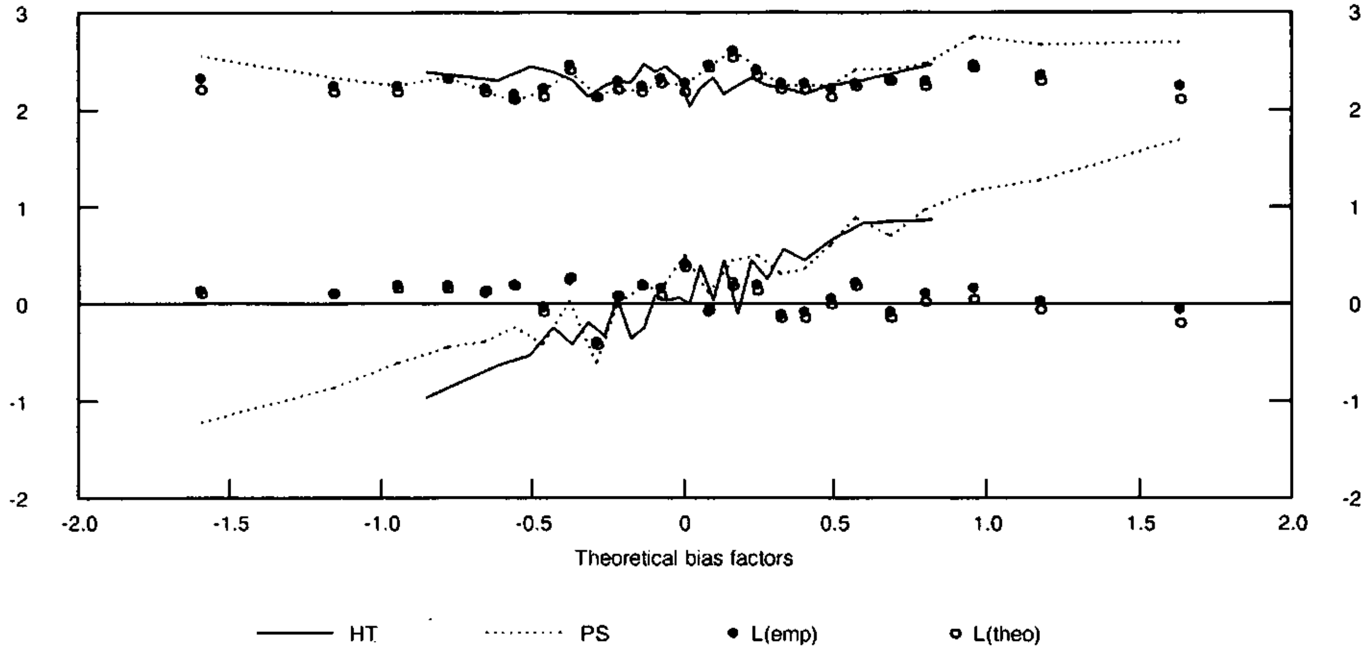

Figure 3. Artificial population simulation, $m = 200$

Table 2 lists unconditional results summarized over all 5,000 samples from each population. Empirical root mean square errors (rmse's) were calculated as $\text{rmse}(\hat{\bar{Y}}) = [\sum_{s=1}^{S}(\hat{\bar{Y}}_s - \bar{Y})^2/S]^{1/2}$ with $S = 5{,}000$ and $\hat{\bar{Y}}_s$ being one of the estimates of the population mean from sample s. In the CPS and artificial populations, results for the Horvitz-Thompson and the ratio estimators were nearly identical so that only the former is shown. Across all samples, the bias of each of the estimators was negligible. As anticipated by the theory, $\hat{\bar{Y}}_{LR}$(theo) was the most precise of the choices, although the largest gain compared to $\hat{\bar{Y}}_{PS}$ was only 4.7% in the artificial population. The need to estimate H destabilizes the regression estimator as shown in the results for $\hat{\bar{Y}}_{LR}$(emp). For the NHIS and CPS populations, $\hat{\bar{Y}}_{LR}$(emp) has a larger root mse than both $\hat{\bar{Y}}_{LR}$(theo) and $\hat{\bar{Y}}_{PS}$. The most noticeable loss is for the NHIS population where the root mse of $\hat{\bar{Y}}_{LR}$(emp) is about 15% larger than that of either $\hat{\bar{Y}}_{LR}$(theo) or $\hat{\bar{Y}}_{PS}$. This result is consistent with the smaller FSU sample size and hence less stable estimate of H for the NHIS population.

Figures 1-3 present conditional simulation results. The 5,000 samples were sorted by the theoretical bias factors presented in section 2.2. The sorting was done separately for each of the estimators of the population mean. In the cases of the two regression estimators, which are theoretically unbiased in large samples, the bias factor for $\hat{\bar{Y}}_{PS}$ was used for sorting. The sorted samples were then put into 25 groups of 200 samples each and empirical biases and root mse's were computed within each group. The group results were then plotted versus theoretical bias factors in the figures. The upper sets of points in each figure are the empirical root mse's of the groups, while the lower sets are empirical biases. The two regression estimators are conditionally unbiased as expected. The other estimators, however, have substantial conditional biases that, in the most extreme sets of samples, are important parts of the mse's. For the CPS population, the range of the bias factors for $\hat{\bar{Y}}_{HT}$ is so much larger (– 10 to 10) than that of the other estimators that we have omitted $\hat{\bar{Y}}_{HT}$ from the plot for clarity. In the neighborhood of the balance point, $\hat{\bar{N}} = \bar{N}$, all estimators perform about the same, but, because of a lack of data at the design stage, we have no control on how close to balance a particular sample may be. The safest choice for controlling conditional bias is, thus, $\hat{\bar{Y}}_{LR}$(emp). This finding is similar to that of Valliant (1990), who noted that, in one-stage, stratified random or systematic sampling, the separate linear regression estimator is a good choice for controlling bias, conditional on the sample mean of an auxiliary variable.

Table 2

Simulation Results for Three Populations. 5,000 Samples were Selected from Each Population

Estimator	Rel-bias $\hat{Y}$ (%)	rmse($\hat{Y}$)	$100*\left[\frac{\text{rmse}(\hat{Y})}{\text{rmse}(\hat{Y}_{PS})} - 1\right]$
	HIS population		
$\hat{Y}_{HT}$	.12	.141	.05
$\hat{Y}_R$	.10	.141	.02
$\hat{Y}_{PS}$	.11	.141	0
$\hat{Y}_{LR}$(emp)	.19	.162	14.71
$\hat{Y}_{LR}$(theo)	.08	.140	−.96
	CPS population		
$\hat{Y}_{HT}$	−.01	10.25	15.8
$\hat{Y}_{PS}$	0	8.85	0
$\hat{Y}_{LR}$(emp)	−.03	9.11	3.0
$\hat{Y}_{LR}$(theo)	−.01	8.79	−.6
	Artificial population		
$\hat{Y}_{HT}$	.02	2.30	−2.93
$\hat{Y}_{PS}$	.12	2.37	0
$\hat{Y}_{LR}$(emp)	.04	2.31	−2.41
$\hat{Y}_{LR}$(theo)	.02	2.26	−4.70

4. DEFECTIVE FRAMES

The conditional biases discussed in the previous sections were of a technical, mathematical nature. A more serious, practical problem in many surveys, that can also lead to bias, is poor coverage of the target population; we address this situation in this section.

4.1 The Basic Problem of Defective Frames

In most real world applications not all of the elementary units in the population are included in the sampling frame. In household surveys, it is not unusual for some demographic subgroups, especially minorities, to be poorly covered by the sampling frame. Bailar (1989), for example, notes that in 1985 the sample estimate from the CPS of the total number of Black males, ages 22-24, was only 73% of an independent estimate of the total population of that group. Corresponding percentages for Black males, ages 25-29 and 60-61, were 80% and 76%.

To formalize the discussion of this type of coverage problem, suppose that $N_{\cdot k}$ now refers to the number of elementary units in the frame and that $\dot{N}_{\cdot k}$ is the *actual* number of population elements in the k^{th} post-stratum. In the discussion below terms with a dot on the top are population values while terms with no dot are frame values. Letting $\dot{Y}_{\cdot k}$ be the aggregate of the y values over all population elements in the k^{th} post-stratum, then it follows that the *true population mean* is given by

$$\dot{\mu} = \lim_{L\to\infty} \frac{\sum_{k=1}^{K} \dot{Y}_{\cdot k}}{\sum_{k=1}^{K} \dot{N}_{\cdot k}} = \lim_{L\to\infty} \sum_{k=1}^{K} \frac{\dot{N}_{\cdot k}}{\dot{N}_{\cdot\cdot}} \frac{\dot{Y}_{\cdot k}}{\dot{N}_{\cdot k}} = \sum_{k=1}^{K} \dot{\phi}_k \dot{\mu}_k .$$

Obviously, all four of the estimators of the mean given in section 2 are biased (both conditionally and unconditionally) for $\dot{\mu}$; the additional bias term being given by $\mu - \dot{\mu}$ for all of the estimators. It should be noted that this bias term is $o(1)$ so it will dominate the other bias terms listed in section 2.2 as the number of FSUs increases. There is another even more basic problem; namely, in most cases the individual frame values $N_{\cdot k}$ are not known so only the ratio estimator is well defined. For example, the Horvitz-Thompson estimator of the mean as defined in section 2 requires $N_{\cdot\cdot}$, the total number of units in the frame, but $N_{\cdot\cdot}$ may be unknown. On the other hand, the $\dot{N}_{\cdot k}$ (or least the proportions $\dot{\phi}_k$) may be known from independent sources and hence be available for the purposes of estimator construction. In household surveys, for instance, the $\dot{N}_{\cdot k}$ may come from intercensal projections of population counts.

Before attempting to construct unbiased estimators for $\dot{\mu}$ it should be noted that

$$\mu - \dot{\mu} = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)(\mu_k - \dot{\mu}_k) + \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\dot{\mu}_k + \sum_{k=1}^{K} \dot{\phi}_k(\mu_k - \dot{\mu}_k) .$$

So, if we assume that for each post-strata the mean of the units in the frame is equal to the true population mean, (*i.e.* $\mu_k = \dot{\mu}_k$ for every k) then the bias term reduces to

$$\mu - \dot{\mu} = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\mu_k = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\dot{\mu}_k .$$

This is very strong (and also very expedient) assumption; however, addressing the problem of defective frame bias without such a condition is virtually impossible.

4.2 Alternative Estimators

The basic strategy is to construct an estimator for the defective frame bias, $\mu - \dot{\mu}$, and then subtract this estimator from the estimators studied earlier. Two cases need to be considered:

Case 1. The frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are unknown, and

Case 2. The frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are known.

Case 1. For this case only the ratio estimator is well defined and the only obvious candidate for an estimator of the bias is

$$\hat{B}_1 = \sum_{k=1}^{K} \left(\frac{\hat{N}_{\cdot k}}{\hat{N}_{\cdot\cdot}} - \dot{\phi}_k \right) \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}} = \hat{\hat{Y}}_R - \sum_{k=1}^{K} \dot{\phi}_k \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}}.$$

Using the strategy given above, the resulting estimator for $\dot{\mu}$ is

$$\hat{\hat{Y}}_1 = \hat{\hat{Y}}_R - \hat{B}_1 = \sum_{k=1}^{K} \dot{\phi}_k \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}}.$$

This is the "post-stratified" estimator usually found in practice. It is straightforward to verify the following properties of $\hat{\hat{Y}}_1$:

$$\mathrm{E}[\hat{\hat{Y}}_1 \mid \hat{\hat{N}}] = \dot{\mu} + [p' P(\hat{\hat{N}} - M_1)] + o(m^{-1})$$

where

$$p' = \left[\frac{\dot{\phi}_1}{\phi_1}, \frac{\dot{\phi}_2}{\phi_2}, \ldots, \frac{\dot{\phi}_K}{\phi_K} \right]$$

$$\mathrm{var}[\hat{\hat{Y}}_1 \mid \hat{\hat{N}}] = m^{-1}[p' V_c p] + o(m^{-(3/2)})$$

$$\mathrm{E}[\hat{\hat{Y}}_1] = \dot{\mu} + o(m^{-1})$$

$$\mathrm{var}[\hat{\hat{Y}}_1] = m^{-1}[p'[\Sigma_{11} - 2D(\mu_k)\Sigma_{21} + D(\mu_k)\Sigma_{22} D(\mu_k)]p] + o(m^{-(3/2)}).$$

The attempt to correct for the defective frame bias is successful in the sense that $\hat{\hat{Y}}_1$ is unconditionally unbiased for $\dot{\mu}$. However, the conditional bias is still present.

Case 2. For this case it can be verified that the estimator

$$\hat{B}_2 = (1 - p)' \left[\hat{\hat{Y}} - \Sigma_{12}\Sigma_{22}^{-1} \left(\frac{\hat{N}}{\hat{N}_{\cdot\cdot}} - M_2 \right) \right],$$

is approximately, conditionally unbiased for $\mu - \dot{\mu}$ and, as $\hat{\hat{Y}}_{LR}$ is conditionally unbiased for μ, it follows directly that the estimator

$$\hat{\hat{Y}}_2 = \hat{\hat{Y}}_{LR} - \hat{B}_2 = p' \left[\hat{\hat{Y}} - \Sigma_{12}\Sigma_{22}^{-1} \left(\frac{\hat{N}}{\hat{N}_{\cdot\cdot}} - M_2 \right) \right]$$

is both conditionally and unconditionally, approximately unbiased for $\dot{\mu}$. It can also be verified that

$$\mathrm{var}[\hat{\hat{Y}}_2 \mid \hat{\hat{N}}] = \mathrm{var}[\hat{\hat{Y}}_2] = m^{-1}[p' V_c p].$$

In addition to the problems of the linear regression estimator cited earlier, this estimator is usually not even well defined as the frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are rarely, if ever, known when the frame is defective.

5. CONCLUSION

This study has generalized the asymptotic techniques suggested by Robinson (1987) to study the problem of post-stratification from a design-based, conditional point-of-view. An important paper in the conditional study of post-stratification was that of Holt and Smith (1979), one of whose basic premises was that $\hat{\hat{Y}}_{PS}$ is conditionally unbiased. This will be true (at least asymptotically) only if $\mathbf{1}'(H - D(\mu_k)) = \mathbf{0}'$; so, in general, this premise is false. In fact, simple random sampling of elementary units may be one of the few realistic cases where this basic premise is true.

From a conditional point of view the linear regression estimator is preferable among the four studied here. Only the regression estimator is conditionally unbiased. The post-stratified estimator is no better (or worse) than either the Horvitz-Thompson or the ratio estimator; all have conditional bias terms of order $m^{-(1/2)}$. All of the estimators have the same conditional variance to terms of order m^{-1}; furthermore, the conditional variance *does not* depend on $\hat{\hat{N}}$, the vector of estimated proportions in the post-strata. Consequently, because of its conditional unbiasedness, the regression estimator has the smallest conditional mean square error.

The Horvitz-Thompson, ratio, and post-stratified estimators are unconditionally unbiased. Although somewhat illogical, one might attempt to make a case for the estimators by comparing their unconditional properties with the conditional properties of the linear regression estimator. But even from this mixed perspective, the $\hat{\hat{Y}}_{LR}$(theo) estimator is clearly superior to the others. Not only is it conditionally unbiased, but the conditional variance of the linear regression estimator can be no larger than the unconditional variance of any of the other estimators. In large FSU samples, the empirical version of the regression estimator will inherit these good properties of $\hat{\hat{Y}}_{LR}$(theo) and also perform well.

The problem of a defectiv: frame introduces complications not found otherwise. Each of the estimators of the mean studied here is biased both conditionally and unconditionally. Bias adjustments are possible only under the restrictive assumption that the mean of units within each post-stratum is the same for all population units whether they are included or excluded from the frame.

An area we have not addressed is variance estimation. A design-based variance estimator for the regression estimator can be obtained using the methods of Särndal, Swensson and Wretman (1989).

ACKNOWLEDGMENT

Any opinions expressed are those of the authors and do not reflect policy of the U.S. Bureau of Labor Statistics. Also, the authors would like to thank the Associate Editor and the referee their germane and constructive comments; we feel they greatly strengthened our paper.

REFERENCES

BAILAR, B. (1989). Information needs, surveys, and measurement errors. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley.

DURBIN, J. (1969). Inferential aspects of randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: Wiley.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society* B, 31, 195-233.

FULLER, W.A. (1981). Comment on an empirical study of the ratio estimator and estimators of its variance by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. 1. New York: John Wiley and Sons.

HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-796.

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

HOLT, D., and SMITH, T.M.F. (1979). Post stratificaticn. *Journal of the Royal Statistical Society A*, 142, 33-46.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association*, 72, 789-827.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LITTLE, R.J.A. (1991). Post-Stratification: A modeler's perspective. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, in press.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

RAO, J.N.K. (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Presented at the Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.

ROYALL, R.M. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart, and Winston.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

VALLIANT, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics*, 6, 115-131.

VALLIANT, R. (1993). Post-stratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3rd. Ed.). London: Griffin.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 193–197
Statistics Canada

Sampling from Imperfect Frames with Unknown Amount of Duplication

SHIBDAS BANDYOPADHYAY and A.K. ADHIKARI[1]

ABSTRACT

This study covers such imperfect frames in which no population unit has been excluded from the frame but an unspecified number of population units may have been included in the list an unspecified number of times each with a separate identification. When the availability of auxiliary information on any unit in the imperfect frame is not assumed, it is established that for estimation of a population ratio or a mean, the mean square errors of estimators based on the imperfect frame are less than those based on the perfect frame for simple random sampling when the sampling fractions of perfect and imperfect frames are the same. For estimation of a population total, however, this is not always true. Also, there are situations in which estimators of a ratio, a mean or a total based on smaller sampling fraction from imperfect frame can have smaller mean square error than those based on a larger sampling fraction from the perfect frame.

KEY WORDS: Imperfect frame; Efficiency.

1. INTRODUCTION

A frequent problem that arises while planning surveys is the non-availability of complete frames. The International Statistical Institute recognized the importance of studying the problem of sampling from imperfect frames and arranged discussions by experts on this topic during its 34th Session held in Ottawa, Canada where Hansen *et al.* (1963) and Szameitat and Schaffer (1963) presented invited papers. One may also refer to Singh (1977, 1983). Wright and Tsao (1983) have written a bibliography on frames to bring attention to problems which arise when sampling from imperfect frames.

Recently two separate surveys were undertaken by the Indian Statistical Institute to evaluate the impact of government sponsored programmes for the uplift of economic conditions of fishermen's community in West Bengal, India. In the first survey (1988), the households were selected using the membership registers of the Fishermen's Co-operative Societies (FCS). In the second and more recent survey, the list of beneficiary fishermen of the Fish Farmer's Development Agency (FFDA) was used. It was known that not all FCS members or FFDA beneficiaries would be from different households, but it was not possible to identify the FCS members or the FFDA beneficiaries belonging to the same household without contacting the households. Thus, when FCS membership registers or FFDA beneficiary lists were used for household selection, the frames contained an unknown number of duplication. Since the household information was collected by personal interview, it was possible to identify the duplication in the selected households only. The values of the variables associated with the households in the sample were divided by the respective number of duplications in the frame while retaining the duplicate households in the sample under separate identification.

The set-up of imperfect frames discussed here is a special case of Rao (1968). One of the referees has pointed out that the situation discussed in the paper also occurs at Statistics Canada in certain frames for business surveys.

Imperfect frames to be covered in this study are those in which no population unit has been excluded from the frame but any population unit may have been included in the frame an unspecified number of times with a separate identification each time. It is assumed that it would be possible to ascertain, at the data collection stage, the number of duplicates in the frame for each selected unit. The possibility of selecting two or more duplicates of a population unit in the sample is not excluded. The availability of auxiliary information on the units in the imperfect frame is not assumed and only simple random sampling without replacement (SRSWOR) schemes are discussed.

Since the total number of population units will not be known from the imperfect frames to be covered here, problems of estimation of a mean of a population character and its total are not identical.

Here is the main question discussed in this paper. Which is better: to up-date the imperfect frame and select a sample, or to use the imperfect frame?

In the two surveys on fishermen's households, it was felt that most of the economic variables of interest would be highly related to the number of FCS members/FFDA beneficiaries in a household in the sense that the variability

[1] Shibdas Bandyopadhyay and A.K. Adhikari, Indian Statistical Institute, Calcutta, India 700 035.

of such an economic variable per FCS member/FFDA beneficiary would be less than the variability of the economic variable per household. It was felt that one could effectively use an imperfect frame in such situations.

It will be established that for situations such as above estimators of a ratio, a mean, or a total based on smaller sampling fraction, imperfect frame can have smaller Mean Square Error (MSE) than those based on a larger sampling fraction from the perfect frame.

Even when the variability is not related to the number of duplications as discussed above, it will be established that for estimating a ratio or a mean, using an imperfect frame will be preferable to using a perfect frame, from the MSE point of view, when the sampling fractions of the imperfect and the perfect frames are same.

2. NOTATIONS AND RELATIONS

Consider a finite population consisting of N units U_1, $U_2, \ldots, U_N$. Let $U_1^*, U_2^*, \ldots, U_M^*$ be the units listed in an imperfect frame. For $k = 1, 2, \ldots, r$, let A_k denote the sub-population of the original N units consisting of N_k distinct population units. Each of the units in A_k is listed in the imperfect frame exactly k number of times under separate identifications. Assume that

(a) each U_i belongs to an A_k for some k, (*i.e.*, each U_i is included in the imperfect frame at least once) and

(b) if U_j^* is selected in the sample using the imperfect frame, it will be possible to identify, at the data collection stage, the corresponding U_i and the associated value of k (*i.e.*, the number of duplicates of U_i in the incomplete frame under separate identifications, one of which is the selected unit U_j^*) for which U_i belongs to A_k.

The following relations are valid.

$$N_1 + N_2 + \ldots + N_r = N;$$

$$N_k \geq 0, k = 1, 2, \ldots, r,$$

$$N_1 + 2N_2 + \ldots + rN_r = M,$$

where $r, N_1, N_2, \ldots, N_r$, and N are all unknown and only M is known with $M \geq N$; M may be written as, for unknown α,

$$M = N(1 + \alpha), \quad \alpha \geq 0. \tag{2.1}$$

Let X and Y values on the unit U_i be X_i and Y_i respectively, $(i = 1, 2, \ldots, N)$. Since each U_j^*, $(j = 1, 2, \ldots, M)$, can be identified with a U_i for some i, $(i = 1, 2, \ldots, N)$, and since U_i belongs to A_k for some k, $(k = 1, 2, \ldots, r)$, define X, Y and C values for the unit U_j^* as

$$X_j^* = X_i/k, \quad Y_j^* = Y_i/k, \quad C_j^* = 1/k.$$

Because of assumptions (a) and (b), X^*, Y^*, and C^* values are observable for the selected units from the imperfect frame.

The following relations connect the measurements in the imperfect frame to those in the perfect frame.

$$\sum_{j=1}^{M} Y_j^* = M\bar{Y}^* = \sum_{i=1}^{N} Y_i = N\bar{Y};$$

$$\sum_{j=1}^{M} C_j^* = M\bar{C}^* = N;$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)^2 = N\sigma_Y^2 - S(2, Y) + (N\bar{Y})^2(1/N - 1/M),$$

where

$$N\sigma_Z^2 = \sum_{i=1}^{N} (Z_i - \bar{Z})^2$$

and

$$S(a,Z) = \sum_{k=2}^{r} (1 - 1/k)\left\{\sum_{i:U_i \in A_k} Z_i^a\right\}; \tag{2.2}$$

$$\sum_{j=1}^{M} (C_j^* - \bar{C}^*)^2 = N(1 - N/M) - S(0, Y);$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)(C_j^* - \bar{C}^*) = N\bar{Y}(1 - N/M) - S(1, Y).$$

For the unit U_i let

$$D_i = Y_i - \bar{Y}; \; W_i = Y_i - RX_i, \quad \text{where} \quad R = \bar{Y}/\bar{X}. \tag{2.3}$$

Since no auxiliary information on the units is assumed, comparisons will be done on the basis of a SRSWOR sample. Let m be the size of the sample from the imperfect frame and n be the corresponding sample size had the frame been perfect. Define efficiency of a perfect frame compared to the corresponding imperfect frame, for any estimator, as

$$\rho = \frac{\text{MSE based on a sample of size } m \text{ from the imperfect frame}}{\text{MSE based on a sample of size } n \text{ had the frame been perfect}}. \tag{2.4}$$

Also define f as the common sampling fraction when the sampling fractions are same, *i.e.*,

$$n = fN, \quad m = fM = n(1 + \alpha). \tag{2.5}$$

3. RESULTS

Before we proceed to answer the main question raised in Section 1 on the choice of sampling from the perfect frame against sampling from the imperfect frame, we briefly look at the alternatives from cost considerations. If the total cost of up-dating the imperfect frame is expected to be more than the additional cost of data collection from the $(m - n)$ extra units, it is economical to use the imperfect frame with a larger sample size than to update the imperfect frame; this is so when

$$\frac{b_1}{b_0}\left(\frac{m - n}{N}\right) \leq 1, \tag{3.1}$$

where b_1 is the per-unit data collection cost and b_0 is the per-unit up-dating cost. It may be noted that one needs to visit effectively N units to up-date the incomplete frame since the remaining $(M - N)$ units are duplicates and can be identified because of assumption (b). It may also be noted that, even from a SRSWOR sample from the imperfect frame, the extra number of units to be canvassed is at most $(m - n)$ since the sample may contain the same unit under separate identifications. These observations lead to (3.1) for preference of using an imperfect frame.

As has been pointed out in Section 1, the total number of population units N will not be known from the imperfect frame. Thus the problems of estimation of a mean and a total are not identical; the problem of estimation of a mean essentially is the problem of estimation of a ratio, but a total can be estimated directly and unbiasedly, based on a SRSWOR sample of size m from the imperfect frame. It is thus appropriate to estimate a population ratio (similar to domain estimation) with estimation of a mean as a special case, and then to treat estimation of a total separately.

3.1 Estimation of a Ratio

For estimation of a ratio $R = (\bar{Y}/\bar{X})$, the usual ratio estimator is

$$\hat{R} = \bar{y}^*/\bar{x}^*,$$

where the lower case letters represent the corresponding quantities based on a sample, $\bar{y}^*$ is the mean of Y^* values based on a sample of size m from the imperfect frame *etc.* $\bar{y}^*$ and $\bar{x}^*$ are respectively unbiased estimators of $(N\bar{Y}/M)$ and $(N\bar{X}/M)$. Using the delta method the MSE of $\hat{R}$, $E(\hat{R} - R)^2$, is given approximately by

$$\frac{M - m}{m(\bar{X}^*)^2(M - 1)M}\sum_{i=1}^{M} W_i^{*2}; \tag{3.2}$$

using the relations of Section 2, (3.2) can be rewritten as

$$\mathrm{MSE}(\hat{R}) = \frac{M(M - m)}{m(N\bar{X})^2(M - 1)}\{N\sigma_W^2 - S(2,W)\},$$

where W values are defined in (2.3) and the W^* values correspondingly obtained. It follows from (2.2) that $S(2,W) \geq 0$, and hence from (3.2) one has

$$0 \leq 1 - \frac{S(2,W)}{N\sigma_W^2} \leq 1. \tag{3.3}$$

It now follows from (2.4) that efficiency ρ is

$$\rho = \frac{nM(M - m)(N - 1)}{mN(N - n)(M - 1)}\left\{1 - \frac{S(2,W)}{N\sigma_W^2}\right\}. \tag{3.4}$$

When sampling fractions are equal, ρ can be written as

$$\rho = \frac{(1 + \alpha)(N - 1)}{(1 + \alpha)(N - 1) + \alpha}\left\{1 - \frac{S(2,W)}{N\sigma_W^2}\right\}. \tag{3.5}$$

It, therefore, follows from (3.3) that ρ given by (3.5) satisfies

$$0 \leq \rho \leq 1 \tag{3.6}$$

and thus it is advantageous to use imperfect frame for estimation of a ratio.

It may be noted that $S(2, W)$ is nondecreasing in α and for fixed α, $S(2, W)$ has a larger value when the units with larger W values are replicated in the imperfect frame. Since σ_W^2 is fixed for a given set of N W values, there may be situations in which ρ in (3.4) is less than 1 (as a matter of fact $S(2, W)$ is equal to $N\sigma_w^2$ when W values are all equal and equal to 0) and consequently, there will be situations when sampling from imperfect frame will be preferable even with smaller sampling fraction to sampling from complete frame.

3.2 Estimation of a Mean

As seen in section 3.1, $\bar{y}^*$ is an unbiased estimator of $(N\bar{Y})/M$ where M is known but N is unknown. Thus it is necessary to estimate N to get an estimator for $\bar{Y}$. It may be noted that $\bar{c}^*$ is an unbiased estimator of (N/M), and thus

$$\hat{\bar{Y}} = \bar{y}^*/\bar{c}^*$$

is a natural ratio-type estimator of $\bar{Y}$. On replacing $\bar{x}^*$ in Section 3.1 by $\bar{c}^*$, the MSE of $\hat{\bar{Y}}$ is given by

$$\text{MSE}(\hat{\bar{Y}}) = \frac{M(M-m)}{mN^2(M-1)} \{N\sigma_D^2 - S(2, D)\},$$

where D values are defined in (2.3). Replacing W in Section 3.1 by D we may conclude that (3.6) holds and imperfect frame is better when (2.5) is true.

3.3 Estimation of a Total

To estimate a total, say $N\bar{Y}$, based on a SRSWOR sample of size m from the imperfect frame, the usual estimator is

$$(\widehat{N\bar{Y}}) = M\bar{y}^*,$$

which is unbiased for $N\bar{Y}$, with variance

$$\text{MSE}(M\bar{y}^*) = \text{Var}(M\bar{y}^*)$$

$$= \frac{M(M-m)}{m(M-1)}$$

$$\left\{N\sigma_Y^2 - S(2,Y) + (N\bar{Y})^2\left(\frac{1}{N} - \frac{1}{M}\right)\right\}.$$

One may write ρ as

$$\rho = \frac{nM(M-m)(N-1)}{mN(N-n)(M-1)}$$

$$\left\{1 - \frac{S(2,Y) - (N\bar{Y})^2(1/N - 1/M)}{N\sigma_Y^2}\right\}.$$

It is clear from the expression of $\text{Var}(M\bar{y}^*)$ that

$$\left\{S(2,Y) - (N\bar{Y})^2\left(\frac{1}{N} - \frac{1}{M}\right)\right\}\Big/ N\sigma_Y^2, \quad (3.7)$$

is less than or equal to unity. However, α and Y values may be so chosen that expression in (3.7) is negative. In such a case, even when (2.5) is true, imperfect frame with larger sampling fraction is inefficient. However, if the scatter of Y^* values are more homogeneous compared to Y values, *i.e.*, if

$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 \geq \sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)^2, \quad (3.8)$$

then the expression in (3.7) is always nonnegative. Now, one can draw similar conclusions as in Section 3.1, for example, (3.6) is valid when (2.5) is true.

4. AN ILLUSTRATION

As pointed out earlier, in the fishermen's survey, ultimate sampling units of beneficiary-fishermen were selected from the list of beneficiaries available. Being a multidisciplinary survey, many characteristics of the sampling units were observed from each of the sampling unit which either related to the household or to the fishing/fishery enterprise to which the sampling unit belonged. Since only the number of beneficiaries (M) was known and the number of corresponding households/enterprises (N) was not known, it was not possible to see the effect of using the imperfect frame for this survey. However for illustration in this paper, we take the samples drawn from one geographical area (a block within an administrative district in the West Bengal State) as our population and see the effect of resampling from it. In this area, there are 27 beneficiaries (M) and 23 distinct enterprises (N), 19 of the enterprises have single ownership (N_1) and 4 are of joint-ownership type (N_2). Our characteristics of interest are the cost of renovation of water areas (Y) and the acreage of operated water areas (X).

The summary statistics of Y and X are as follows:

$$\sum Y_i = 58{,}815, \quad \sum X_i = 23.36,$$

$$R = \left(\sum Y_i\right)\Big/\left(\sum X_i\right) = 2{,}517.77,$$

$$S(2,Y) = 212{,}201{,}800, \quad S(2,D) = 145{,}101{,}018,$$

$$S(2,W) = 104{,}505{,}327,$$

$$23\sigma_Y^2 = 442{,}702{,}791, \quad 23\sigma_X^2 = 13.6503 \quad \text{and}$$

$$23\sigma_W^2 = 394{,}790{,}716,$$

where W is defined in (2.3).

To find the effect of sampling from the list of 27 beneficiaries we find estimates of

R = Renovation cost per acre of water area,
$\bar{X}$ = Average water area per enterprise in acre and
$N\bar{X}$ = Total acreage of water areas operated by all 23 enterprises.

The table below gives the efficiencies for different choices of m and n.

Efficiency of sampling from perfect frame compared to sampling from imperfect frame (ρ)

Sample sizes		Efficiency for estimators of		
n	m	R	$\bar{X}$	$N\bar{X}$
2	2	0.8695	0.6453	0.9508
4	4	0.8841	0.6561	0.9668
6	6	0.9022	0.6696	0.9866
8	8	0.9225	0.6866	1.0117
8	9	0.7791	0.5781	0.8519
10	10	0.9551	0.7088	1.0444
10	11	0.8172	0.6065	0.8937

It can be seen that in most cases sampling from imperfect frame are more efficient.

ACKNOWLEDGEMENT

Authors wish to thank an Associate Editor and the referees for their valuable suggestions towards improvement of this paper.

REFERENCES

HANSEN, M.H., HURWITZ, W.N., and JABINE, T.N. (1963). The Use of imperfect lists for probability sampling at the U.S. Bureau of Census. *Bulletin of the International Statistical Institute*, 40, 497-517, (with discussions).

INDIAN STATISTICAL INSTITUTE (1988). *A study of Fishermen in West Bengal*: 1985-1986.

RAO, J.N.K. (1968). Some non-response sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

SINGH, R. (1977). A note on the use of incomplete multi-auxiliary information in sample surveys. *Australian Journal of Statistics*, 19, 105-107.

SINGH, R. (1983). On the use of incomplete frames in sample surveys. *Biometrical Journal*, 25, 545-549.

SZAMEITAT, K., and SCHAFFER, K.A. (1963). Imperfect frames in statistics and the consequences for their use in sampling. *Bulletin of the International Statistical Institute*, 40, 517-538, (with discussions).

WRIGHT, T., and TSAO, H.J. (1983). *A frame on frames: An annonated bibliography. Statistical Methods and Improvement of Data Quality*, (Ed. T. Wright). New York: Academic Press, 25-72.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 199–204
Statistics Canada

An Alternative View of Forest Sampling

FRANCIS A. ROESCH, JR., EDWIN J. GREEN and CHARLES T. SCOTT[1]

ABSTRACT

A generalized concept is presented for all of the commonly used methods of forest sampling. The concept views the forest as a two-dimensional picture which is cut up into pieces like a jigsaw puzzle, with the pieces defined by the individual selection probabilities of the trees in the forest. This concept results in a finite number of independently selected sample units, in contrast to every other generalized conceptualization of forest sampling presented to date.

KEY WORDS: Forest sampling; PPS sampling.

1. INTRODUCTION

The sampling of forests is often accomplished as a two part process: first a random point is located in the forest and then a cluster of trees in the vicinity of the point is selected for the sample by some rule. The two most common rules are known as (circular, fixed-area) plot sampling and (horizontal) point sampling. In the former, all trees for which the center of the cross-section of the bole at 4.5 feet above the ground is within a constant horizontal distance (d) of the random point are included in the sample. In the latter, tree i is selected for the sample if this center is within a horizontal distance αr_i of the random point, where r_i is the radius of the cross-section and α is a constant, chosen appropriately to obtain a desired sampling intensity. Tree i would be selected with probability proportional to πd^2 in plot sampling (the probability is the same for all trees) and with probability proportional to πr_i^2 (basal area of tree i) in point sampling (larger trees have a higher probability of selection).

There has been much discussion in the forestry literature about what the sample unit actually is in the various methods of forest sampling. The tree is considered the sample unit from one point of view (*e.g.* Oderwald 1981), while from other points of view, the cluster of trees associated with the point (*e.g.* Palley and Horwitz 1961; Schreuder 1970), the circular plot (*e.g.* Cunia 1965), and the point (*e.g.* Husch 1955) are considered the sample units. These various viewpoints are supported by different statistical tools. For example, treating the tree as the sample unit requires the use of finite population sampling theory, while considering the point as sample unit requires the use of the somewhat more advanced theory of infinite population sampling. In addition, plot sampling has traditionally been presented from the viewpoint of the plot as the sample unit, whereas point sampling has usually been presented from the viewpoints of the tree or the point as the sample unit. Therefore, these very common and quite similar sampling mechanisms artificially appear disparate.

We will show a conceptualization of the primary sample unit that is applicable to every type of forest sampling scheme which selects trees based on the location of a random point. We will also show that this conceptualization is simple and that it provides a finite number of mutually exclusive and independently selected sample units. This is in contrast to the view of the tree or the cluster of trees as the sample unit, because trees are not selected independently and clusters of trees are not mutually exclusive. It also differs from the views of the randomly placed point or the plot as the sample unit, because there are an infinite number of units in these cases. We will also suggest that this alternative conceptualization is often more appropriate.

2. THE JIGSAW PUZZLE VIEW

Suppose that there are N trees in the forest with labels $1, 2, \ldots, N$. Associated with the N trees are values of interest $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N\}$, K-circles $K = \{K_1, K_2, \ldots, K_N\}$, and selection areas of sizes $\tilde{A} = \{\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_N\}$. Grosenbaugh and Stover (1957) first defined the K-circle in the context of point sampling. For our purposes the K-circle of tree i, K_i, is an imaginary circle, centered at tree center, with radius d in plot sampling and radius αr_i in point sampling. The selection area for tree i, of size A_i (in acres), is the portion of tree i's K-circle which is within the forest, and is the area from within which a random point will select the tree for the sample.

When discussing point sampling, Palley and Horwitz (1961) contend that "... the primary sampling unit is a cluster of trees associated with a locus of origin. The locus

[1] Francis A. Roesch, Jr., Mathematical Statistician, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, 701 Loyola Avenue, New Orleans, LA 70113; Edwin J. Green, Professor of Forestry, Cook College-Rutgers University, P.O. Box 231, New Brunswick, NJ 08903; and Charles T. Scott, Project Leader, Forest Ecosystem Modeling Unit, Northeastern Forest Experiment Station, USDA Forest Service, 359 Main Road, Delaware, Ohio 43015.

of origin is a point in the case of point sampling . . .". Actually the locus of origin is not a point because the cluster of trees is not selected only from that point but rather from an infinite set of points within a specific area.

We offer the alternative view of the sample units being the mutually exclusive sections of ground resulting from the overlapping selection areas of the individual trees in the forest.

The treatment of the ground broken up into primary sampling units is clearly shown in Figure 1, for example. The correspondence between the population, sampling frame and sample unit as given in say Cochran (1977, p. 6) is apparent: the population (*or the puzzle picture*) is divided up into mutually exclusive, exhaustive sample units (*the puzzle pieces*) which together comprise the sample frame. Each ground segment has a definite probability of selection and the total of these probabilities over all segments is 1. We will call this the jigsaw puzzle view.

Associated with each ground segment are attributes of interest, the measurement of which will result in identical values from any point in that segment of ground. The crux of the matter is that individual points are equivalent within any particular segment. The ground segments, of course, are selected with probability proportional to size. In the case of point sampling, the segment size is determined by the basal areas and spatial distribution of the trees and the constant α chosen. Once α is chosen, the sample frame at a particular point in time is fixed. In the case of plot sampling, the size of the segment is determined by d and the spatial distribution of the trees. Thus, regardless of the method used to determine the sample trees (*e.g.*, plot sampling or point sampling), all schemes can be thought of as cutting the puzzle up in some way, selecting the pieces with probability proportional to their size, and then turning each piece over to read the attributes associated with it.

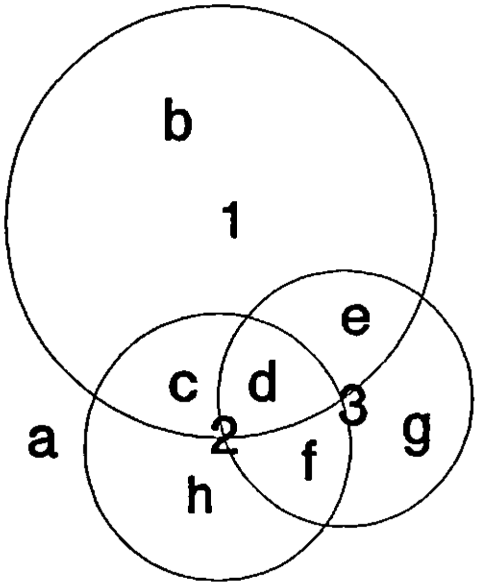

Figure 1. The Puzzle Pieces. Trees 1, 2 and 3 are centered at their respective numbers. The surrounding circles represent the selection areas of the trees. Each of the lettered segments represents a sample unit.

Returning to our proposition that this view is often more appropriate, we note that the purpose of most forest surveys is to describe the *forest*, not the individual trees. Our aggregations are usually made on a per acre or hectare basis, *i.e.* units of the forest land, not units of the tree. From the same place we may measure many other things besides the trees such as topographic and site characteristics. It is therefore usually more appropriate to view pieces of the forest as the sample units rather than individual trees in the forest.

Although we will be working mostly in the context of forest sampling in general, our discussion is easily applied to any specific type of forest sampling which relies on the selection of trees by some function of randomly placed points. The only difference is the definition of the ground segments, or how we dissect the picture into puzzle pieces. For example, in plot sampling the ground is divided into pieces defined by overlapping circles of equal size, while in point sampling the definition is by overlapping circles of sizes proportional to each corresponding tree's basal area.

To examine this further, suppose that we randomly drop a point on the surface of a forest and use any function to select sample trees. Suppose also that within our forest are three trees (1, 2, and 3) whose selection areas overlap. In Figure 1, trees 1, 2 and 3 are centered at their respective numbers with their selection areas shown as circles. Each lettered segment represents a different sample unit. If the point falls in segment a, the empty cluster is chosen, in segment b, the cluster containing only tree 1, in segment d, the cluster of all three trees, *etc.* Tree 1 would therefore be selected from segments b, c, d or e. This results in a situation somewhat analogous to that described in Kish (1965, sec. 11.2), if we were to consider the tree to be the primary sample unit, in which a list to be sampled from contains duplicate listings of the same unit. In this case, the list would be one of clusters of trees, in which most trees are associated with more than one cluster. The clusters are selected with probability proportional to the size of the ground segment. The standard technique of weighting duplicate elements of a list, discussed by Kish, considers rather the selection of primary units with equal probability.

The jigsaw puzzle view reduces the complexity of the sampling mechanism in one sense by first mapping the tree population into the ground segment population and thereby reducing the sample list from a list of clusters of trees in which trees belong to more than one cluster to a list of unique ground segments. Our claim below that forest sampling simulations can be simplified by the jigsaw puzzle view is supported wholly by the tradeoff between the one time cost of this reduction in the complexity of the sample list and the need to select from that list many times.

To map the tree population into the segment population, an observation for a segment would preferably be the sum of weighted tree values, the weight for each tree being proportional to its probability of being observed from that particular segment. The probability that sampled tree i was selected from the particular ground segment j is:

$$p_{ij} = \left(\frac{A_j}{\tilde{A}_i}\right) Z_{ij},$$

where:

A_j = the area of segment j in acres, and

$$Z_{ij} = \begin{cases} 1 \text{ if segment } j \text{ is part of the } k\text{-circle of tree } i \\ 0 \text{ otherwise.} \end{cases}$$

The sum over j of p_{ij} is 1. We can now write the observation for each segment as a sum of weighted tree values:

$$y_j = \sum_{i=1}^{N} p_{ij} \tilde{y}_i. \qquad (1)$$

Now suppose that we randomly drop m points on the surface of a forest with the same assumptions as above (our sampling is with replacement). An unbiased estimator of the total value of interest for a sample selected with probability proportional to size is:

$$\begin{aligned} \hat{Y} &= \frac{A_T}{m} \sum_{j=1}^{m} \frac{y_j}{A_j} \\ &= \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j, \end{aligned} \qquad (2)$$

where:

$A_T = \sum_{j=1}^{M} A_j$; the total area of the forest in acres,

m = the number of sample points,

M = the number of ground segments, and

W_j = the number of times the jth unit appears in the sample.

Note that W_j is an integer between 0 and m, inclusive. A_j and y_j are fixed and W_j is random. In addition, we will define:

$Y = \sum_{i=1}^{N} \tilde{y}_i$; the total value of interest across all trees, and

$Y^* = \sum_{j=1}^{M} y_j$; the total value of interest across all segments.

To show that $\hat{Y}$ is unbiased for Y, we will first show $\hat{Y}$ to be unbiased for Y^* and then show that Y^* equals Y. Following Cochran (1977, p. 252-255), we can show $\hat{Y}$ to be unbiased for Y^*:

$$\begin{aligned} E[\hat{Y}] &= E\left[\frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j\right] \\ &= \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} E[W_j]. \end{aligned} \qquad (3)$$

W_j is a multinomial random variable and its expected value is equal to $m(A_j/A_T)$. Therefore

$$E[\hat{Y}] = \sum_{j=1}^{M} y_j = Y^*. \qquad (4)$$

We can now show that $\hat{Y}$ is unbiased for Y by showing that $Y^* = Y$. Substituting the right hand side of equation (1) for y_j in the definition of Y^*, we get:

$$Y^* = \sum_{j=1}^{M} \sum_{i=1}^{N} p_{ij} \tilde{y}_i. \qquad (5)$$

After substituting in the definition of p_{ij} and rearranging the order of summation:

$$Y^* = \sum_{i=1}^{N} \tilde{y}_i \left[\frac{1}{\tilde{A}_i} \sum_{j=1}^{M} A_j Z_{ij}\right]. \qquad (6)$$

Because

$$\tilde{A}_i = \sum_{j=1}^{M} A_j Z_{ij},$$

the term within the brackets on the right hand side of (6) equals 1, and

$$Y^* = \sum_{i=1}^{N} \tilde{y}_i = Y. \quad \text{Q.E.D.} \qquad (7)$$

By definition, the variance of $\hat{Y}$ is

$$V(\hat{Y}) = \left(\frac{1}{mA_T}\right) \sum_{j=1}^{M} A_j \left(\frac{A_T y_j}{A_j} - Y\right)^2. \qquad (8)$$

The sample estimate of the variance is then (Cochran 1977):

$$v(\hat{Y}) = \frac{1}{m(m-1)} \sum_{j=1}^{m} \left(\frac{A_T y_j}{A_j} - \hat{Y} \right)^2 . \qquad (9)$$

The general development in equations (1) through (9) can be used for any specific type of forest sampling which follows the two part process of selecting trees from randomly placed points.

As a further example of the use of the jigsaw puzzle view, we will illustrate the sample frame when point samples are used to measure forest growth. For the greatest efficiency, measurements are taken at two points in time and the same random points are used both times. This type of sampling for forest growth is known as remeasured point sampling and has been discussed at length in the literature, most recently by Van Deusen *et al.* (1986) and Roesch *et al.* (1989, 1991, 1993). If a remeasured point sample had been taken, and Figure 1 represented time 1, the puzzle for the overall sample might be cut up into pieces like those in Figure 2. Trees 1, 2 and 3 are the same as those in Figure 1 and tree 4 is a tree which grew into the stand between times 1 and 2. The inner circles represent the trees' point sample areas of selection at time 1 (say αr_{i1}, including a subscript for time) and the outer circles represent the point sample areas of selection at time 2 (αr_{i2} is larger due to an increase in basal area). Tree 4 only has an outer circle since it did not exist at time 1 and tree 2 only has an inner circle since it died prior to time 2. The dotted circle represents the selection area tree 2 would have had at time 2 if time 2 had occurred just prior to the tree's demise. Therefore, the dotted circle does not contribute to the definition of the segments.

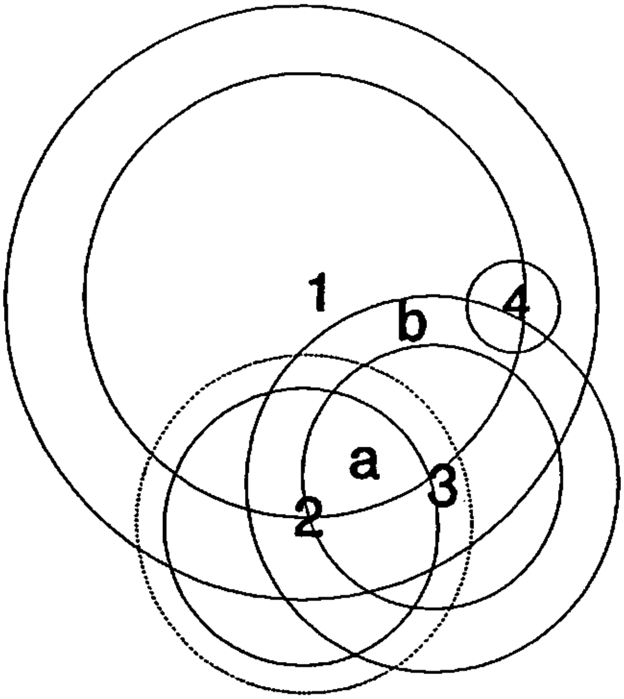

Figure 2. Puzzle pieces defined by location, size, and time. An example of sample units in a remeasured point sample. Trees 1 and 3 have grown and survived, tree 2 grew somewhat before dying and tree 4 is ingrowth.

If the random point lands in segment *a*, trees 1 and 3 would be measured at both times and tree 2 would be measured only at time 1; in segment *b*, tree 1 would be measured at both times and tree 3 would only be measured at time 2. This exemplifies the fact that even though another dimension was added to the sample (the time dimension), the forest sample concept remains the same, since the time dimension can be collapsed down onto the puzzle picture. So, in addition to the conditions mentioned above, the definition of the segments depends upon the exact times of each measurement. This concept of the sample unit is helpful in understanding the estimators of the components of change from time 1 to time 2 given in Van Deusen *et al.* (1986) and Roesch *et al.* (1989 and 1991).

3. DISCUSSION

Given the simplicity of the jigsaw puzzle concept, one might wonder why this view of forest sampling has not been proposed before. The most compelling reason is probably that the above estimators cannot be calculated when the A_j's are unknown. Since a particular tree's area of selection might be divided between many of the puzzle pieces and the size of a particular puzzle piece may be limited by trees not sampled by that piece, the selection areas of both sample and non-sample trees must be known to calculate the A_j's of the selected segments. For example, referring to Figure 1, if our point landed in section *c*, we would sample trees 1 and 2 and the area of $c + d$ would be readily calculable. However, to calculate $\hat{Y}$ and $v(\hat{Y})$, we need the area of *c* alone, for which we do not have adequate sample information. We will show that this apparent deficiency is unimportant by showing that $\hat{Y}$ can be reexpressed in terms which are calculable. This will, in fact, always be the case no matter which sampling method is described by the jigsaw puzzle view.

The jigsaw puzzle view of point sampling is actually a mapping of the tree population into the associated ground segment population. We can reexpress $\hat{Y}$ to show that it is equivalent to the usual point sampling estimator which is based upon the tree population. Expanding equation (2) to include the definition of y_j and subsequent rearrangement gives:

$$\begin{aligned}
\hat{Y} &= \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j \\
&= \frac{A_T}{m} \sum_{j=1}^{M} \frac{\sum_{i=1}^{N} p_{ij} \tilde{y}_i}{A_j} W_j \\
&= \frac{A_T}{m} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\tilde{y}_i Z_{ij} W_j}{\tilde{A}_i} \qquad (10) \\
&= \frac{A_T}{m} \sum_{i=1}^{N} \frac{\tilde{y}_i}{\tilde{A}_i} \sum_{j=1}^{M} Z_{ij} W_j \\
&= \frac{A_T}{m} \sum_{i=1}^{N} \frac{\tilde{y}_i}{\tilde{A}_i} w_i,
\end{aligned}$$

where w_i equals the number of times tree i is selected for the sample. The final expression in (10) is the usual point sample estimator.

The purpose of this paper, therefore, is not to introduce a new set of estimators for sampling systems which already have reasonably good estimators, but rather to show how sampling schemes of quite disparate justifications in the literature are related in general. This alternative avenue of understanding may be useful in many ways. For one, we believe that some abstract forest sampling systems may be easier to understand if put into the framework described above. Our experience is that students, for instance, readily grasp the idea of point sampling when taught as merely a method of dividing the forest up into non-overlapping jigsaw puzzle pieces which are then sampled with probability proportional to size. Researchers who are interested in developing new forest sampling schemes or new estimators for existing schemes may benefit from this view because it provides another path for understanding new sampling schemes and for programming the forest sampling simulations used to test the new methods. The simulation discussed in Roesch (1993), for example, was simplified by using the jigsaw puzzle view rather than the other conceptualizations of the forest sampling frame which had been suggested up to that time. The simplification stemmed from the fact that the bulk of the simulation could be used for many different sampling schemes with only minor modifications to the subroutine which dissected the puzzle.

Because forest sampling simulations often start with a mapped forest, the A_j's are readily obtainable. Once the puzzle is dissected, y_j can be calculated for each piece. The simulator then simply selects these pieces from a list in proportion to their size. Contrast this with the simulation resulting from the view of the point as the sample unit. In this latter simulation, a random point would be dropped and the tree list searched for all of the trees close enough to that point to be selected for the sample. Then the attributes of interest would be calculated. Since the probability of selecting a point from an infinite population twice is zero, this list search and calculation would have to be repeated for each random point, possibly resulting in repeated calculation of the attributes from the same cluster of trees. For simulation purposes, the optimal approach to programming will depend upon the length of the tree list to be searched, the degree of clustering in the tree population, and the number of random points.

4. CONCLUSION

We've presented a generalized forest sampling concept which utilizes a finite number of ground segments as the sample units existing within a land-area based sample frame. We have also given estimators based on this concept. The jigsaw puzzle view should be of help in understanding the similarities and differences between different methods of forest sampling by putting all of the methods into the same framework. Although we would not normally utilize the associated estimators in their given form in an actual forest survey, we can always find an equivalent calculable form. The additional benefit of an alternative route for sampling simulations is not only one of academics but also economics. Given the amount of time and money it takes to acquire data in forestry studies, the ability to easily test the properties of different sampling methods before they are applied in the field is of paramount importance. We would not endeavor to undermine the importance of a thorough theoretical development of proposed forest sampling schemes as the crucial first step, but simulation of these schemes before implementation may help uncover overlooked problems. This alternative conceptualization will, in general, facilitate comparisons within any group of forest sampling schemes.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor, two anonymous referees and Hans Schreuder for their helpful comments.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

CUNIA, T. (1965). Continuous forest inventory, partial replacement of samples and multiple regression. *Forest Science*, 11, 480-502.

GROSENBAUGH, L.R., and STOVER, W.S. (1957). Point-sampling compared with plot-sampling in southeast Texas. *Forest Science*, 3, 2-14.

HUSCH, B. (1955). Results of an investigation of the variable plot method of cruising. *Journal of Forestry*, 53, 570-574.

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

ODERWALD, R.G. (1981). Point and plot sampling - the relationship. *Journal of Forestry*, 79, 377-378.

PALLEY, M.N., and HORWITZ, L.G. (1961). Properties of some random and systematic point sampling estimators. *Forest Science*, 7, 52-65.

ROESCH, F.A. Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Science*, 39. In press.

ROESCH, F.A. Jr., GREEN,E.J., and SCOTT, C.T. (1989). New compatible estimators for survivor growth and ingrowth from remeasured horizontal point samples. *Forest Science*, 35, 281-293.

ROESCH, F.A. Jr., GREEN, E.J., and SCOTT C.T. (1991). Compatible basal area and number of trees estimators from remeasured horizontal point samples. *Forest Science*, 37, 136-145.

ROESCH, F.A. Jr., GREEN, E.J., and SCOTT, C.T. (1993). A test of alternative estimators for volume at time 1 from remeasured point samples. *Canadian Journal of Forest Research*, 23, 598-604.

SCHREUDER, H.T. (1970). Point sampling theory in the framework of equal-probability cluster sampling. *Forest Science*, 16, 240-246.

VAN DEUSEN, P.C., DELL, T.R., and THOMAS, C.E. (1986). Volume growth estimation from permanent horizontal points. *Forest Science*, 32, 415-422.

Survey Methodology, December 1993
Vol. 19, No. 2, pp. 205–215
Statistics Canada

Panel Surveys: Adding the Fourth Dimension

GRAHAM KALTON and CONSTANCE F. CITRO[1]

ABSTRACT

Surveys across time can serve many objectives. The first half of the paper reviews the abilities of alternative survey designs across time - repeated surveys, panel surveys, rotating panel surveys and split panel surveys - to meet these objectives. The second half concentrates on panel surveys. It discusses the decisions that need to be made in designing a panel survey, the problems of wave nonresponse, time-in-sample bias and the seam effect, and some methods for the longitudinal analysis of panel survey data.

KEY WORDS: Panel surveys; Rotating panel surveys; Repeated surveys; Panel attrition; Time-in-sample bias; Seam effect; Longitudinal analysis.

1. INTRODUCTION

Survey populations are constantly changing over time, both in composition and in the characteristics of their members. Changes in composition occur when members enter the survey population through birth (or reaching adulthood), immigration, or leaving an institution (for a noninstitutional population) or leave through death, emigration, or entering an institution. Changes in characteristics include, for example, a change from married to divorced, or from a monthly income of $2,000 to one of $2,500. These population changes give rise to a range of objectives for the analysis of survey data across time. This paper reviews survey designs that produce the data needed to satisfy these various objectives.

The paper is divided into two parts. The first part contains a review of the general issues involved in conducting surveys across time, including the objectives of such surveys and the types of survey design that may be employed. This part is to be found in Section 2. The second, and main, part of the paper discusses one particular survey design, a panel survey that follows the same sample of units through time. The considerations involved in designing, conducting, and analyzing a panel survey are reviewed in Section 3. Section 4 provides some concluding remarks.

2. SURVEYS ACROSS TIME

This section presents an overview of analytic objectives across time, of designs for surveys across time, and of the extent to which different designs can satisfy the various objectives. The discussion relies heavily on Duncan and Kalton (1987), which contains a more detailed treatment of these issues.

Changes in population characteristics and composition over time lead to a variety of objectives for surveys across time. These objectives include the following:

(a) The estimation of population parameters (*e.g.*, the proportion of the population in poverty) at distinct time points.

(b) The estimation of average values of population parameters across time (*e.g.*, the daily intake of iron averaged across a year).

(c) The estimation of net changes, that is changes at the aggregate level (*e.g.*, the change in the proportion of unemployed from one month to the next).

(d) The estimation of gross changes and other components of individual change (*e.g.*, the proportion of persons who were in poverty one year and were not in poverty in the following year).

(e) The aggregation of data for individuals over time (*e.g.*, the summation of twelve monthly incomes to give annual income).

(f) The collection of data on events occurring in a specified time period (*e.g.*, becoming unemployed), and on their characteristics (*e.g.*, duration of spells of unemployment).

(g) The cumulation of samples over time, especially samples of rare populations (*e.g.*, women who become widowed).

(h) The maintenance of a sample of members of a rare population that was identified at one point of time (*e.g.*, scientists and engineers identified from a large-scale survey at one point of time).

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850; Constance F. Citro, National Research Council, 2101 Constitution Ave. N.W., Washington, D.C., U.S.A., 20418.

A number of survey designs have been developed to provide the data needed to address these objectives. These designs are:

- *Repeated survey.* A repeated survey is a series of separate cross-sectional surveys conducted at different time points. No attempt is made to ensure that any of the same elements are sampled for the individual surveys. The elements are sampled from a population defined in the same manner for each individual survey (*e.g.*, the same geographical boundaries and age-limits) and many of the same questions are asked in each survey.
- *Panel survey.* A panel survey collects the survey data for the same sample elements at different points of time.
- *Repeated panel survey.* A repeated panel survey is made up of a series of panel surveys each of a fixed duration. There may be no overlap in the time period covered by the individual panels, for instance one panel may start only as (or after) the previous one ends, or there may be an overlap, with two or more panels covering part of the same time period.
- *Rotating panel survey.* Strictly, a rotating panel survey is equivalent to a repeated panel survey with overlap. Both limit the length of a panel, and have two or more panels in the field at the same time. However, it seems useful to distinguish between the two designs because they have different objectives. Rotating panel surveys are widely used to provide a series of cross-sectional estimates and estimates of net change (*e.g.*, of unemployment rates and changes in such rates), whereas repeated panel surveys with overlaps also have a major focus on longitudinal measures (*e.g.*, durations of spells of unemployment). In consequence, repeated panel surveys tend to have longer durations and have fewer panels in operation at any given time than rotating panel surveys.
- *Overlapping survey.* Like a repeated survey, an overlapping survey is a series of cross-sectional surveys conducted at different time points. However, whereas the repeated survey does not attempt to secure any sample overlap from the survey at one time point to the next, an overlapping survey is designed to provide such overlap. The aim may be to maximize the degree of sample overlap while taking into account both the changes desired in selection probabilities for sample elements that remain in the survey population and also changes in population composition over time.
- *Split panel survey.* A split panel survey is a combination of a panel survey and a repeated survey or rotating panel survey.

The choice of design in a particular case depends on the objectives to be satisfied. Some designs are better than others for some objectives but poorer for other objectives. Some designs cannot satisfy certain objectives at all. For a detailed discussion, see Duncan and Kalton (1987).

The strength of a repeated survey is that it selects a new sample at each time point, so that each cross-sectional survey is based on a probability sample of the population existing at that time. A panel survey is based on a sample drawn from the population existing at the start of the panel. Although attempts are sometimes made to add samples of new entrants to a panel at later time points, such updating is generally difficult to do and is done imperfectly. Moreover, nonresponse losses from a panel as it ages heighten concerns about nonresponse bias when the panel sample is used to estimate cross-sectional parameters for later time points. For these reasons, repeated surveys are stronger than panel surveys for producing cross-sectional and average cross-sectional estimates (objectives (a) and (b)). With average cross-sectional estimates, another factor to be considered is the correlation between the values of the survey variables for the same individual at different time points. When this correlation is positive, as it generally is, it increases the standard errors of the average cross-sectional estimates from a panel survey. This factor thus also favours repeated surveys over panel surveys for average cross-sectional estimates.

The superior representation of the samples for a repeated survey at later time points also argues in favour of a repeated survey over a panel survey for estimating net change (assuming that the interest in net change relates to changes in both population composition and characteristics). However, in this case the positive correlations of the values of the survey variables for the same individuals across time decreases the standard errors of estimates of net change from a panel survey. Hence the presence of this correlation operates in favour of the panel design for measuring net change.

The key advantages of the panel design are its abilities to measure gross change, and also to aggregate data for individuals over time (objectives (d) and (e)). Repeated surveys are incapable of satisfying these objectives. The great analytic potential provided by the measurement of individual changes is the major reason for using a panel design.

Repeated surveys can collect data on events occurring in a specified period and on durations of events (*e.g.*, spells of sickness) by retrospective questioning. However, retrospective questioning often introduces a serious problem of response error in recalling dates, and the risk of telescoping bias. A panel survey that uses a reference period for the event that corresponds to the interval between waves of data collection can eliminate the telescoping problem by using the previous interview to bound the recall (*i.e.*, an illness reported at the current interview can be discarded if it had already been reported at the previous one). Similarly, a panel survey can determine the duration of an event from successive waves of data collection, limiting the length of recall to the interval between waves.

Repeated data collections over time can provide a vehicle for accumulating a sample of members of a rare population, such as persons with a rare chronic disease or persons who have recently experienced a bereavement. Repeated surveys can be used in this manner to generate a sample of any form of rare population. Panel surveys, however, can be used to accumulate only samples of new rare events (such as bereavements) not of stable rare characteristics (such as having a chronic disease). If a sample of members with a rare stable characteristic (*e.g.*, persons with doctoral degrees) has already been identified, a panel survey can be useful for maintaining the sample over time, with suitable supplementation for new entrants at later waves (for an example, see Citro and Kalton 1989).

Rotating panel surveys are primarily concerned with estimating current levels and net change (objectives (a) and (c)). As such, elements are usually retained in the panel for only short periods. For instance, sample members remain in the monthly Canadian Labour Force Survey for only six months. The extent to which individual changes can be charted and aggregation over time can be performed is thus limited by the short panel duration. A special feature of rotating panel surveys is the potential to use composite estimation to improve the precision of both cross-sectional estimates and estimates of net change (see Binder and Hidiroglou 1988; Cantwell and Ernst 1993). See also Fuller *et al.* (1993) for an alternative method of using past information in forming estimates from a rotating panel design.

Like rotating panel surveys, overlapping surveys are primarily concerned with estimating current levels and net change. They can also provide some limited information on gross change and aggregations over time. Overlapping survey designs are applicable in situations where some sample overlap is required and where the desired element selection probabilities vary over time. This situation arises in particular in establishment surveys, where the desired selection probability for an establishment may vary from one cross-sectional survey to the next to reflect its change in size and type of activity. In such circumstances, a Keyfitz-type procedure can be applied to maximize the retention of elements from the previous survey while taking account of changes in selection probabilities and population composition (see, for example, Keyfitz 1951; Kish and Scott 1971; Sunter 1986). The U.S. Internal Revenue Service Statistics of Income Division's corporate sample provides an example of an overlapping survey design (Hinkins *et al.* 1988).

By combining a panel survey with a repeated survey or a rotating panel survey, a split panel survey can provide the advantages of each. However, given a constraint on total resources, the sample size for each component is necessarily smaller than if only one component had been used. In particular, estimates of gross change and other measures of individual change from a split panel survey will be based on a smaller sample than would have been the case if all the resources had been devoted to the panel component.

In comparing alternative designs for surveys across time, the costs of the designs need to be considered. For instance, panel surveys avoid the costs of repeated sample selections incurred with repeated surveys, but they face costs of tracking and tracing mobile sample members and sometimes costs of incentives to encourage panel members to continue to cooperate in the panel (see Section 3). If two designs can each satisfy the survey objectives, the relative costs for given levels of precision for the survey estimates need to be examined.

3. PANEL SURVEYS

The repeated measures over time on the same sampled elements that are obtained in panel surveys provide such surveys with a key analytic advantage over repeated surveys. The measurements of gross change and other components of individual change that are possible with panel survey data form the basis of a much greater understanding of social processes than can be obtained from a series of independent cross-sectional snapshots. The power of longitudinal data derived from panel surveys has long been recognized (see, for instance, Lazarsfeld and Fiske 1938; Lazarsfeld 1948), and panel surveys have been carried out in many fields for many years. Subjects of panel surveys have included, for example, human growth and development, juvenile delinquency, drug use, victimizations from crime, voting behaviour, marketing studies of consumer expenditures, education and career choices, retirement, health, and medical care expenditures. (See Wall and Williams (1970) for a review of early panel studies on human growth and development, Boruch and Pearson (1988) for descriptions of some U.S. panel surveys, and the Subcommittee on Federal Longitudinal Surveys (1986) for descriptions of U.S. federal panel surveys.) In recent years, there has been a major upsurge in interest in panel surveys in many subject-matter areas, and especially in household economics. The ongoing U.S. Panel Study of Income Dynamics began in 1968 (see Hill 1992 for a description of the PSID) and similar long-term panel studies have been started in the past decade in many European countries. The U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) in 1983 (Nelson *et al.* 1985; Kasprzyk 1988; Jabine *et al.* 1990), and Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) in 1993. The growth in interest in panel surveys has also given rise to an increase in literature about the methodology of such surveys, including such recent texts as Kasprzyk *et al.* (1989), Magnusson and Bergman (1990), and Van de Pol (1989).

This section reviews the major issues involved in the design and analysis of panel surveys. The treatment is geared towards repeated panel surveys of fixed duration like the SIPP and SLID, but most of the discussion applies more generally to all forms of panel survey.

3.1 Design Decisions for a Panel Survey

The time dimension adds an extra dimension of complexity to a panel survey as compared with a cross-sectional survey. In addition to all the decisions that need to be made about the design features of a cross-sectional survey, a wide range of extra decisions needs to be reached for a panel survey. Major design decisions include:

- *Length of the panel.* The longer the panel lasts, the greater is the wealth of data obtained for longitudinal analysis. For instance, the longer the panel, the greater the number of spells of unemployment starting during the life of the panel that will be completed before the end of the panel, and hence the greater the precision in estimating the survival function for such spells. On the other hand, the longer the panel, the greater the problems of maintaining a representative cross-sectional sample at later waves, because of both sample attrition and difficulties in updating the sample for new entrants to the population.

  It can sometimes be beneficial to vary the length of the panel between different types of panel members. Thus, for instance, when the analytic objectives call for it, panel members with certain characteristics (*e.g.*, members of a minority population) or who experience certain events during the course of the regular panel (*e.g.*, a divorce) can be retained in the panel for extended periods of observation.

- *Length of the reference period.* The frequency of data collection depends on the ability of respondents to recall the information collected in the survey over time. Thus, the PSID, with annual waves of data collection, requires recall of events occurring in the previous calendar year, whereas SIPP, with four-monthly waves of data collection, requires recall for the preceding four months. The longer the reference period, the greater the risk of recall error.

- *Number of waves.* In most cases the number of waves of data collection is determined by a combination of the length of the panel and the length of the reference period. The greater the number of waves, the greater the risk of panel attrition and time-in-sample effects, and the greater the degree of respondent burden.

- *Overlapping or non-overlapping panels.* With a repeated panel survey of fixed duration, a decision needs to be made as to whether the panels should overlap across time. Consider, for instance, the proposal of a National Research Council study panel that the SIPP should be a four-year panel (Citro and Kalton 1993). One possibility is to run each panel for four years, starting a new panel when the previous one finishes. Another possibility is run each panel for four years, but starting a new panel every two years. Yet another possibility is to run each panel for four years, starting a new panel every year.

  The design of nonoverlapping panels has the benefit of simplicity, since only one panel is in the field at any one time. It also produces a large sample for longitudinal analysis; for instance, the panels with the nonoverlapping design can be roughly twice the size of those with the design that has two overlapping panels at any one time. However, this increase in sample size for nonoverlapping panels does not apply for cross-sectional estimates, since the data from the panels covering a given time point can be combined for cross-sectional estimation. Also, the cross-sectional estimates for a time period near the end of a panel with the nonoverlapping design are at greater risk of bias from attrition, time-in-sample bias, and failure to update the sample fully for new population entrants than is the case with an overlapping design, in which one panel is of more recent origin. Moreover, the overlapping design permits the examination of such biases through a comparison of the results for the two panels for a given time period, whereas no such examination is possible with a nonoverlapping design. Another limitation of the nonoverlapping design is that it may not be well positioned to measure the effect of such events as a change in legislation. For instance, if legislation takes effect in the final year of a nonoverlapping panel, there will be little opportunity to evaluate its effect by comparing the situations of the same individuals before and for some period after the legislation is enacted. With overlapping panels, one of the panels will provide a wider window of observation.

- *Panel sample size.* For a given amount of annual resources, the sample size for each panel is determined by the preceding factors. A larger panel for longitudinal analysis can be achieved by lengthening the reference period and by employing a nonoverlapping design. The sample size for cross-sectional estimates can be increased by lengthening the reference period, but not by using a nonoverlapping design.

The above list determines the major parameters of a panel survey design, but there still remain a number of other factors that need to be considered:

- *Mode of data collection.* As with any survey, a decision needs to be made as to whether the survey data are to be collected by face-to-face interviewing, by telephone, or by self-completion questionnaire, and whether computer assisted interviewing (CAPI or CATI) is to be used. With a panel survey, this decision needs to be made for each wave of data collection, with the possibility of different modes for different waves (for instance, face-to-face interviewing at the first wave to make contact and establish rapport, with telephone interviewing or mail questionnaires at some of the later waves). When modes may be changed between waves, consideration needs to be given to the comparability of the data across waves. Sometimes a change in mode may involve a change in interviewer, as for instance would occur with a change from face-to-face interviewing to a centralized CATI operation. Then the effects of a change of interviewer between waves on the respondent's willingness to continue in the panel and on the comparability of responses across waves also need to be carefully considered.

- *Dependent interviewing*. With panel surveys there is the possibility of feeding back to respondents their responses at earlier waves of data collection. This dependent interviewing procedure can secure more consistent responses across waves, but risks generating an undue level of consistency. The ease of application of dependent interviewing depends on the length of the interval between waves and the mode of data collection. Processing the responses from one wave to feed back in the next is easier to accomplish if the interval between waves is a long one and if computer assisted interviewing is employed. Edwards *et al.* (1993) describe the use of dependent interviewing with CAPI in the Medical Care Beneficiary Survey, a survey which involves three interviews per year with each respondent.
- *Incentives*. Monetary or other incentives (*e.g.*, coffee mugs, calculators, lunch bags) may be offered to sampled persons to encourage their participation in a survey. With a panel survey, incentives may be used not only to secure initial participation but also to maintain cooperation throughout the duration of the panel. There is an issue of when are the best times to provide incentives in a panel survey (*e.g.*, at the first wave, at an intermediate wave, or at the last wave of the panel). Panel survey researchers often send respondents a survey newsletter, frequently giving some recent highlights from the survey findings, at regular intervals, both to generate goodwill for the survey and to maintain contact with respondents (see below). Birthday cards sent at the time of the respondents' birthdays are also often used for these purposes.
- *Respondent rules*. Survey data are often collected from proxy informants when respondents are unavailable for interview. With a panel survey, this gives rise to the possibility that the data may be collected from different individuals at different waves, thus jeopardizing the comparability of the data across waves. The respondent rules for a panel survey need to take this factor into account.
- *Sample design*. The longitudinal nature of a panel survey needs to be considered in constructing the sample design for the initial wave. Clustered samples are commonly employed for cross-sectional surveys with face-to-face interviewing in order to reduce fieldwork travel costs and to enable frame construction of housing unit listings to be performed only for selected segments. These benefits are bought at the price of the increase in the variance of survey estimates arising from the clustering. The optimum extent of clustering depends on the various cost factors involved and the homogeneity of the survey variables in the clusters (see, for instance, Kish 1965). With a panel survey, the use and extent of any clustering should be determined in relation to the overall panel with all its waves of data collection. In particular, the benefit of reduced fieldwork costs disappears for waves of data collection that are conducted by telephone interviewing or mail questionnaire. Also the migration of panel members to locations outside the original clusters reduces the benefit of the initial clustering for fieldwork costs at later waves. (However, some benefits of the initial clustering still operate for the large proportion of mobile persons who move within their own neighbourhoods.)

Oversampling of certain population subgroups is widely used in cross-sectional surveys to provide sufficient numbers of subgroup members for separate analysis. Such subgroups may, for instance, comprise persons with low incomes, minority populations, persons in a specified age-group, or persons living in certain geographical areas. Such oversampling can also be useful in panel surveys, but caution is needed in its application. With long-term panels, one reason for caution is that the objectives of the survey may change over time. Oversampling to meet an objective identified at the start of a panel may prove harmful to objectives that emerge later. Another reason for caution is that many of the subgroups of interest are transient in nature (*e.g.*, low income persons, persons living in a given geographical area). Oversampling persons in such subgroups at the outset of the panel may be of limited value for later waves: some of those oversampled will leave the subgroup while others not oversampled will join it. Thirdly, the definition of the desired subgroup for longitudinal analysis needs to be considered. For instance, SIPP data are used to estimate durations of spells on various welfare programs. Since such estimates are usually based on new spells starting during the life of the panel, it may not be useful to oversample persons already enrolled on welfare programs. See Citro and Kalton (1993) for a discussion of oversampling for the SIPP.

When oversampling of a certain subgroup of the population (*e.g.*, a minority population) is desired for a panel survey, the oversampling may require a large screening operation. The assessment of the cost of such screening should be made in the context of the full panel with all its waves of data collection. An expensive screening operation at the first wave may well be justifiable in this context.

- *Updating the sample.* When the sole objective of a panel survey is longitudinal analysis, it may be sufficient to adopt a cohort approach that simply follows the initial sample selected for the first wave. However, when cross-sectional estimates are also of interest, it may be necessary to update the sample at each wave to represent new entrants to the population. Updating for all types of new entrants is often difficult, but it is sometimes possible to develop fairly simple procedures to account for certain types of new entrants. For instance, in a panel of persons of all ages, babies born to women panel members after the start of the panel can be included as panel members. The SIPP population of inference comprises persons aged 15 and over. By identifying in initial sampled households persons who are under 15 years old but who will attain that age before the end of the panel, by following them during the panel, and by interviewing them after they reach 15 years of age, a SIPP panel can be updated for this class of new entrants (Kalton and Lepkowski 1985).

  Attention also needs to be paid to panel members who leave the survey population. For some the departure is clearly permanent (*e.g.*, deaths), but for others it may be only temporary (*e.g.*, going abroad or entering an institution). If efforts are made to keep track of temporary leavers, they can be readmitted to the panel if they return to the survey's population of inference.

  Panel surveys such as SIPP and PSID collect data not only for persons in original sampled households, but also for other persons - nonsampled persons - with whom they are living at later waves. The prime purpose of collecting survey data for nonsampled persons is to be able to describe the economic and social circumstances of sampled persons. The issue arises as to whether any or all nonsampled persons should remain in the panel after they stop living with sampled persons. For some kinds of analysis it is useful to follow them. However, to follow them would eat significantly into the survey's resources.

  When data are collected for nonsample members, these data may be used simply to describe the circumstances of sample members, in which case analyses are restricted to sample members, with nonsample members being assigned weights of zero. Alternatively, nonsample members can be included in cross-sectional analyses. In this case appropriate weights for sample and nonsample persons need to be developed to reflect the multiple ways in which individuals may appear in the dataset. Huang (1984), Ernst (1989) and Lavallée and Hunter (1993) describe the fair share weighting approach that may be used for this purpose.

- *Tracking and tracing.* Most panel surveys encounter the problem that some panel members have moved since the last wave and cannot be located. There are two ways to try to handle this problem. First, attempts can be made to avoid the problem by implementing procedures for tracking panel members between waves. One widely-used procedure when there is a long interval between waves is to send mailings, such as birthday cards and survey newsletters, to respondents between waves, requesting the post office to provide notification of change of address if applicable. Another tracking device is to ask respondents for the names and addresses or telephone numbers of persons close to them (*e.g.*, parents) who are unlikely to move and who will be able to provide locating information for them if they move.

  The second way to deal with lost panel members is to institute various tracing methods to try to locate them. With effort and ingenuity, high success rates can be achieved. Some methods of tracing may be specific for the particular population of interest (*e.g.*, professional societies for persons with professional qualifications) while others may be more general, such as telephone directories, computerized telephone number look-ups, reverse telephone directories for telephone numbers of neighbours, mail forwarding, marriage licence registers, motor vehicle registrations, employers, and credit bureaus. It can be useful to search death records for lost panel members, particularly for long-term panel surveys. Panel members found to have died can then be correctly classified, rather than being viewed as non-respondents. Methods of tracing are discussed by Burgess (1989), Clarridge *et al.* (1978), Crider *et al.* (1971) and Eckland (1968).

3.2 Problems of Panel Surveys

Panel surveys share with all surveys a wide range of sources of nonsampling error. This section does not review all these sources, but rather concentrates on three sources that are unique to panel surveys, namely wave nonresponse, time-in-sample bias and the seam effect.

3.2.1 Wave nonresponse

The nonresponse experienced by panel surveys at the first wave of data collection corresponds to that experienced by cross-sectional surveys. The distinctive feature of panel surveys is that they encounter further nonresponse at subsequent waves. Some panel members who become non-respondents at a particular wave do not respond at any subsequent wave while others respond at some or all subsequent waves. The former are often termed attrition cases and the latter non-attrition cases. The overall wave nonresponse rates in panel surveys increase with later waves, but with well-managed surveys the rate of increase usually declines appreciably over time. For example, with the 1987 SIPP panel, the sample loss was 6.7% at wave 1, 12.6% at wave 2, and it then increased slowly to 19.0% at wave 7 (Jabine *et al.* 1990). The tendency for the nonresponse rate to flatten off at later waves is comforting, but nevertheless the accumulation of nonresponse over many waves produces high nonresponse rates at later waves of a long-term panel. For instance, in 1988, after 21 annual rounds of data collection, the PSID nonresponse rate for individuals who lived in 1968 sampled households had risen to 43.9% (Hill 1992).

The choice between the two standard general-purpose methods for handling missing survey data - weighting adjustments and imputation - is not straightforward for wave nonresponse in panel surveys. For longitudinal analysis, the weighting approach drops all records with one or more missing waves from the data file and attempts to compensate for them by weighting adjustments applied to the remaining records. This approach can lead to the loss of a substantial amount of data when the data file covers several waves. On the other hand, the imputation approach retains all the reported data, but requires conducting wholesale imputations for missing waves. A compromise approach uses imputation for some patterns of wave nonresponse (*e.g.*, those with only one missing wave, where data are available from both adjacent waves), and weighting for others (see, for example, Singh *et al.* 1990). For cross-sectional analysis, separate data files may be created for each wave. These files can comprise all the respondents for that wave, with either weighting adjustments or imputations for the wave nonrespondents. Kalton (1986) and Lepkowski (1989) discuss general methods for handling wave nonresponse, Lepkowski *et al.* (1993) discuss imputations for wave nonresponse in the SIPP, and Michaud and Hunter (1993) describe plans for handling wave nonresponse in the SLID.

With wave nonresponse there is the possibility of collecting some or all of the data for the missing wave at a subsequent interview. However, the quality of the retrospective data collected in this way needs to be carefully assessed. An experiment was conducted to examine the utility of this approach with the 1984 SIPP panel, using a missing wave form to collect responses for a skeleton set of core questions for the missing wave (Huggins 1987; Singh 1993). The analyses showed substantially fewer transitions in receipt of income, assets, and government assistance from the missing wave form than from benchmark data. In consequence the use of the missing wave form was discontinued. Administrative records may sometimes provide another possible source of skeletal data for missing waves.

3.2.2 Time-in-sample bias

Time-in-sample bias, or panel conditioning, refers to the effect that panel members' responses at a given wave of data collection are affected by their participation in previous waves. The effect may reflect simply a change in reporting behaviour. For example, a respondent may recognize from previous interviews that a "Yes" response to a question leads to follow-up questions, whereas a "No" answer does not. The respondent may therefore give a "No" answer to avoid the burden of the extra questions. Alternatively, a respondent may learn from previous interviews that detailed information on income is needed, and may therefore prepare for later interviews by collecting the necessary data. The time-in-sample effect may also reflect a change in actual behaviour. For example, a respondent may enroll in the food stamp program as a result of learning of its existence from the questions asked about it at earlier waves of data collection.

A recent experimental study of panel conditioning in a four-year panel study of newlyweds found some evidence that participation in the study did affect marital well-being (Veroff *et al.* 1992). However, that study used in-depth interviewing techniques that are more intrusive than those used in most surveys. A number of studies of panel conditioning that have been conducted in more standard survey settings have found that conditioning effects do sometimes occur, but they are not pervasive (Traugott and Katosh 1979; Ferber 1964; Mooney 1962; Waterton and Lievesley 1989).

A benefit of rotating and overlapping panel surveys is that they enable estimates for the same time period obtained from different panels to be compared. Such comparisons have clearly identified the presence of what is termed "rotation group bias" in the U.S. and Canadian Labour Force Surveys (*e.g.* Bailar 1975, 1989, and U.S. Bureau of the Census 1978, for the U.S. Current Population Survey; Ghangurde 1982, for the Canadian Labour Force Survey). Rotation group bias may reflect nonresponse bias and conditioning effects. In analyses comparing the overlapping 1985, 1986 and 1987 SIPP panels, Pennell and Lepkowski (1992) found few differences in the results from the different panels.

3.2.3 Seam effect

Many panel surveys collect data for subperiods within the reference period from the last wave of data collection. The SIPP, for instance, collects data on a monthly basis within the four-month reference period between waves. The seam effect refers to the common finding with this form of data collection that the levels of reported changes between adjacent subperiods (*e.g.*, going on or off of a welfare program from one month to the next) are much greater when the data for the pair of subperiods are collected in different waves than when they are collected in the same wave. The seam effect has been found to be pervasive in SIPP, and to relate to both recipiency status and amounts received (see, for example, Jabine *et al.* 1990; Kalton and Miller 1991). It has also been found in PSID (Hill 1987). Murray *et al.* (1991) describe approaches used to reduce the seam effect in the Canadian Labour Market Activity Survey.

3.3 Longitudinal Analysis

There is a substantial and rapidly expanding literature on the analysis of longitudinal data, including a number of texts on the subject (*e.g.* Goldstein 1979; Hsiao 1986; Kessler and Greenberg 1981; Markus 1979). This treatment cannot be comprehensive, but rather identifies a few general themes.

- *Measurement of gross change.* As has already been noted, a key analytic advantage of a panel survey over a repeated survey is the ability to measure gross change, that is, change at the individual level. The basic approach to measuring gross change is the turnover table that tabulates responses at one wave against the responses to the same question at another wave. The severe limitation to this form of analysis is that changes in measurement errors across waves can lead to serious bias in the estimation of the gross change (for further discussion, see Kalton *et al.* 1989; Rodgers 1989; Abowd and Zellner 1985; Chua and Fuller 1987; Fuller 1990; and Skinner 1993).

- *Relationship between variables across time.* Panel surveys collect the data necessary to study the relationships between variables measured at different times. For instance, based on the data collected in the 1946 British birth cohort, the National Survey of Health and Development, Douglas (1975) found that children who were hospitalized for more than a week or who had repeated hospitalizations between the ages of 6 months and 3½ years exhibited more troublesome behaviour in school and lower reading scores at age 15. In principle, cross-section surveys may employ retrospective questions to collect the data needed to perform this type of analysis. However, the responses to such questions are often subject to serious memory error, and potentially to systematic distortions that affect the relationships investigated.

- *Regression with change scores.* Regression with change scores can be used to avoid a certain type of model misspecification. Suppose that the correct regression model for individual i at time t is

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it},$$

where x_{it} is an explanatory variable that changes value over time and z_{it} is an explanatory variable that is constant over time (*e.g.*, gender, race). Suppose further that z_{it} is unobserved; it may well be unknown. Then β can still be estimated from the regression on the change scores:

$$Y_{i(t+1)} - Y_{(t)} = \beta(x_{i(t+1)} - x_{it}) + \epsilon_{i(t+1)} - \epsilon_{it},$$

(Rodgers 1989; Duncan and Kalton 1987).

- *Estimation of spell durations.* The data collected in panel surveys may be used to estimate the distribution of lengths of spells of such events as being on a welfare program. In panel surveys like the SIPP, some individuals have a spell in progress at the start of the panel (initial-censored spells), some start a spell during the panel, and some spells continue beyond the end of the panel (right-censored spells). Thus, not all spells are observed in their entirety. The distribution of spell durations may be estimated by applying survival analysis methods, such as the Kaplan-Meier product-limit estimation procedure to all new spells (including right-censored new spells) starting during the life of the panel (*e.g.* Ruggles and Williams 1989).

- *Structural equation models with measurement errors.* The sequence of data collection in a panel survey provides a clear ordering of the survey variables that fits well with the use of structural equation modelling for their analysis. This form of analysis can make allowance for measurement errors, and with several repeated measures can handle correlated error structures (*e.g.* Jöreskog and Sörbom 1979).

4. CONCLUDING REMARKS

The data sets generated from panel surveys are usually extremely rich in analytic potential. They contain repeated measures for some variables that are collected on several occasions, and also measures for other variables that are asked on a single wave. Repeated interviewing of the same sample provides the opportunity to collect data on new variables at each wave, thus yielding data on an extensive range of variables over a number of waves. A panel data set may be analyzed both longitudinally and cross-sectionally. Repeated measures may be used to examine individual response patterns over time, and they may also be related to other variables. Variables measured at a single wave may be analyzed both in relation to other variables measured at that wave and to variables measured at other waves.

The richness of panel data is of value only to the extent that the data set is analyzed, and analyzed in a timely manner. Running a panel survey is like being on a treadmill: the operations of questionnaire design, data collection, processing and analysis have to be undertaken repeatedly for each successive wave. There is a real danger that the survey team will become overwhelmed by this process, with the result that the data are not fully analyzed. To avoid this danger, adequate staffing is needed and a well-integrated organization needs to be established.

In addition it is advisable to keep the panel survey design simple. The survey design should be developed to meet clearly-specified objectives. Adding complexities to the design to enhance the richness of the panel data set for other uses should be critically assessed. Although persuasive arguments can often be made for such additions, they should be rejected if they threaten the orderly conduct of any stage of the survey process.

As noted earlier, measurement errors have particularly harmful effects on the analysis of individual changes from panel survey data. The allocation of part of a panel survey's resources to measure the magnitude of such errors is therefore well warranted (Fuller 1989). Measurement errors may be investigated either by validity studies (comparing survey responses with "true" values from an external source) or by reliability studies (*e.g.*, reinterview studies). The results of such studies may be then used in the survey estimation procedures to adjust for the effects of measurement errors.

REFERENCES

ABOWD, H.M., and ZELLNER, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 1-24.

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

BORUCH, R.F., and PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.

BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 52-74.

CANTWELL, P.J., and ERNST, L.R. (1993). New developments in composite estimation for the Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 121-130.

CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

CITRO, C.F., and KALTON, G. (1989). *Surveying the Nation's Scientists and Engineers*. Washington DC: National Academy Press.

CITRO, C.F., and KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington DC: National Academy Press.

CLARRIDGE, B.R., SHEEHY, L.L., and HAUSER, T.S. (1978). Tracing members of a panel: a 17-year follow-up. *Sociological Methodology*, (Ed. K.F. Schuessler). San Francisco: Jossey-Bass, 389-437.

CRIDER, D.M., WILLITS, F.K., and BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.

DOUGLAS, J.W.B. (1975). Early hospital admissions and later disturbances of behaviour and learning. *Developmental Medicine and Child Neurology*, 17, 456-480.

DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

ECKLAND, B.K. (1968). Retrieving mobile cases in longitudinal surveys. *Public Opinion Quarterly*, 32, 51-64.

EDWARDS, W.S., SPERRY, S., and EDWARDS, B. (1993). Using CAPI in a longitudinal survey: a report from the Medicare Current Beneficiary Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 21-30.

ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh), New York: John Wiley, 139-159.

FERBER, R. (1964). Does a panel operation increase the reliability of survey data: the case of consumer savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.

FULLER, W.A. (1989). Estimation of cross-sectional and change parameters: Discussion. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 480-485.

FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1993). Estimators for longitudinal surveys with application to the U.S. Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 309-324.

GHANGURDE, P.D. (1982). Rotation group bias in the LFS estimates. *Survey Methodology*, 8, 86-101.

GOLDSTEIN, H. (1979). *The Design and Analysis of Longitudinal Studies*. New York: Academic Press.

HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

HINKINS, S., JONES, H., and SCHEUREN, F. (1988). Design modifications for the SOI corporate sample: Balancing multiple objectives. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-221.

HSIAO, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.

HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

HUGGINS, V. (1987). Evaluation of missing wage data from the Survey of Income and Program Participation (SIPP). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 205-209.

JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation Quality Profile.* Bureau of the Census, Washington DC: U.S. Department of Commerce.

JÖRESKOG, K.G., and SÖRBOM, D. (1979). *Advances in Factor Analysis and Structural Equation Models.* Lanham MD: University Press of America.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., KASPRZYK, D., and McMILLEN, D.B. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 249-270.

KALTON G., and LEPKOWSKI, J.M. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.

KALTON, G., and MILLER, M.E. (1991). The seam effect with Social Security income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.

KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues.* SIPP Working Paper No. 8830. Washington DC: U.S. Bureau of the Census.

KASPRZYK, D., DUNCAN G., KALTON, G., and SINGH, M.P. (Eds.) (1989). *Panel Surveys.* New York: John Wiley.

KESSLER, R.C., and GREENBERG, D.F. (1981). *Linear Panel Analysis.* New York: Academic Press.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46, 183-201.

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

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

LAVALLÉE, P., and HUNTER, L. (1993). Weighting for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 65-75.

LAZARSFELD, P.F. (1948). The use of panels in social research. *Proceedings of the American Philosophical Society*, 42, 405-410.

LAZARSFELD, P.F., and FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.

LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 348-374.

LEPKOWSKI, J.M., MILLER, D.P., KALTON, G., and SINGH, R. (1993). Imputation for wave nonresponse in the SIPP. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 99-109.

MAGNUSSON, D., and BERGMAN, L.R. (Eds.) (1990). *Data Quality in Longitudinal Research.* New York: Cambridge University Press.

MARKUS, G.B. (1979). *Analyzing Panel Data.* Beverly Hills, CA: Sage Publications.

MICHAUD, S., and HUNTER, L. (1993). Strategy for minimizing the impact of non-response for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 89-98.

MOONEY, H.W. (1962). *Methodology in Two California Health Surveys.* Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.

MURRAY, T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference.* U.S. Bureau of the Census. Washington DC: U.S. Department of Commerce, 715-730.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1.* SIPP Working Paper No. 8401. Washington DC: U.S. Bureau of the Census.

PENNELL, S.G., and LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.

RODGERS, W.L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 432-456.

SINGH, R.P. (1993). Methodological experiments in the Survey of Income and Program Participation. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 157-166.

SINGH, R., HUGGINS, V., and KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys.* SIPP Working Paper No. 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.

SKINNER, C.J. (1993). Logistic modelling of longitudinal survey data with measurement error. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 269-276.

SUBCOMMITTEE ON FEDERAL LONGITUDINAL SURVEYS (1986). *Federal Longitudinal Surveys.* Statistical Policy Working Paper 13. Washington DC: Office of Management and Budget.

SUNTER, A.B. (1986). Implicit longitudinal files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

TRAUGOTT, M., and KATOSH, K. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43, 359-377.

U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey Design and Methodology.* Bureau of the Census Technical Paper No. 40, Washington DC: U.S. Government Printing Office.

VAN DE POL, F.J.R. (1989). *Issues of Design and Analysis of Panels.* Amsterdam: Sociometric Research Foundation.

VEROFF, J., HATCHETT, S., and DOUVAN, E. (1992). Consequences of participating in a longitudinal study of marriage. *Public Opinion Quarterly*, 56, 315-327.

WALL, W.D., and WILLIAMS, H.L. (1970). *Longitudinal Studies and the Social Sciences.* London: Heinemann.

WATERTON, J., and LIEVESLEY, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 319-339.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1993:

C.H. Alexander, *U.S. Bureau of the Census*
M.G. Arellano, *Advanced Linkage Technologies of America*
J. Armstrong, *Statistics Canada*
T.S. Arthanari, *Indian Statistical Institute*
M. Bankier, *Statistics Canada*
Y. Beaucage, *Statistics Canada*
T.R. Belin, *University of California - Los Angeles*
W. Bell, *U.S. Bureau of the Census*
D. Bellhouse, *University of Western Ontario*
E. Berumen, *IASI*
J. Bethel, *Westat*
D.A. Binder, *Statistics Canada*
G.J. Brackstone, *Statistics Canada*
J.M. Brick, *Westat*
P.A. Chollette, *Statistics Canada*
G.H. Choudhry, *Statistics Canada*
A. Chu, *Westat*
M.J. Colledge, *Statistics Canada*
L. Cox, *U.S. Environmental Protection Agency*
J.-C. Deville, *INSEE*
P. Dick, *Statistics Canada*
J.D. Drew, *Statistics Canada*
R.E. Fay, *U.S. Bureau of the Census*
G. Forsman, *University of Linkoping*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistics Canada*
M.E. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
M.A. Hidiroglou, *Statistics Canada*
H. Hogan, *U.S. Bureau of the Census*
G.J.C. Hole, *Statistics Canada*
D. Holt, *University of Southampton*
D.G. Horvitz, *Retired*
D.A. Hubble, *U.S. Bureau of the Census*
A.Z. Israëls, *Netherlands Central Bureau of Statistics*
A.E. Johnson, *Agency for Health Care Policy and Research*
G. Kalton, *Westat*
P.S. Kott, *U.S. Department of Agriculture*
P. Lahiri, *University of Nebraska - Lincoln*
D. Lalande, *Statistics Canada*
P. Lavallée, *Statistics Canada*
L. Lebart, *CNRS - Télécom Paris*
H. Lee, *Statistics Canada*
R.J.A. Little, *University of California - Los Angeles*
J. Liu, *Research Triangle Institute*
D. Malec, *National Centers for Health Statistics*
J.T. Massey, *U.S. Department of Health and Human Services*
S.M. Miller, *U.S. Bureau of Labor Statistics*
W.J. Mitofsky, *Voter Research and Surveys*
G. Nathan, *Hebrew University*
H.B. Newcombe, *Consultant*
C.A. Patrick, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistics Canada*
E. Rancourt, *Statistics Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
L. Roy, *Statistics Canada*
D. Royce, *Statistics Canada*
D.B. Rubin, *Harvard University*
K. Rust, *Westat*
I. Sande, *Bell Communications Research*
C.-E. Särndal, *Université de Montréal*
O. Sautory, *INSEE*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
H.T. Schreuder, *U.S. Department of Agriculture*
J. Sedransk, *State University of New York - Albany*
C. Skinner, *University of Southampton*
A. Singh, *Statistics Canada*
T.M.F. Smith, *University of Southampton*
N.L. Spruill, *U.S. Office of the Secretary of Defence*
K.P. Srinath, *Statistics Canada*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
S. Sudman, *University of Illinois - Urbana-Champaign*
J.-L. Tambay, *Statistics Canada*
M.E. Thompson, *University of Waterloo*
A.R. Tupek, *U.S. Bureau of Labor Statistics*
J. Waksberg, *Westat*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*
J. Wretman, *Stockholm University*
A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1993 issues: S. Beauchamp and S. Lineger (Photocomposition), and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M.M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 3, 1993

Printed in Great Britain at the Alden Press, Oxford

Special Issue
Confidentiality and Data Access
JOS 1993:2

Recent years have seen the unfortunate marriage of two issues of great concern to statistical agencies: confidentiality protection and steadily increasing nonresponse rates. To respond to these growing concerns, the Panel on Confidentiality and Data Access of the Committee on National Statistics and the Journal of Official Statistics have produced this special issue on Confidentiality and Data Access.

Contents

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:

I. 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.