SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2000

•

VOLUME 26

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2000 • VOLUME 26 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

February 2001

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology 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

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household 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 Survey Methodology (Catalogue no. 12-001-XPB) is CDN $47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 x 2 issues); Other Countries, CDN $20 ($10 x 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

LESLIE KISH
(1910 - 2000)

This issue is dedicated to the memory of Leslie Kish. His infectious joie de vivre, his deep concern for the oppressed and the underprivileged, and his profound contributions to survey methodology and statistics have been, and continue to be, an inspiration to so many.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 26, Number 2, December 2000

CONTENTS

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 117-118
Statistics Canada

In This Issue

This issue is dedicated to Leslie Kish, who passed away this fall at the age of 90. It is remarkable to note that to the end of his life Professor Kish continued to propose and develop new ideas in statistics and survey methodology, as evidenced by his article "Cumulating/Combining Population Surveys" which appeared one year ago in the 25th anniversary issue of this journal. This issue of *Survey Methodology* opens with a reflection on his life and contributions to statistics written by Ivan Fellegi.

The paper by Haines, Pollock and Pantula examines the estimates of a total when two incomplete list frames are combined with an area frame. The authors give suggestions on appropriate population totals to account for the incompleteness of the frames. In addition, their models allow for the fact that larger sampling units are more likely to be included on the incomplete list frames.

Beaumont proposes an estimation method which reduces the bias induced by a response mechanism that depends on the variable of interest, known as a nonignorable response mechanism. The proposed method requires one model for the variable of interest and one model for the response probability. The method is considered robust with respect to the hypothesis of normality since it is constructed in such a way that there is no need to specify the error distribution of model involving the variable of interest, unlike the method of maximum likelihood. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

Spencer considers the problem of estimating the design effect due to weighting when the selection probabilities are correlated with the variable of interest. Using a regression representation of the population, Spencer presents an approximation to the design effect when the selection probabilities are correlated with the variable of interest.

Biemer and Bushery use the Markov assumption on labour force transitions to identify classification errors in labour force data. Using this methodology, they estimate response error rates in panels of monthly labour force data from the Current Population Survey (CPS). The general consistency of the results is taken as an indicator that Markov Latent Class Analysis is a useful method to assess the accuracy of responses in the CPS. Critical to this analysis is confirming the Markov assumption; the authors present some interesting empirical evidence for its validity over the short term in the CPS.

Many statistical offices use modified half-sample-replication (MHS) for estimating the sampling variance of medians. This is an important practical problem because direct calculation of sample medians can be computationally intensive. An alternative estimation method is to group the continuous data into discrete intervals and use linear interpolation over the interval containing the median. In their paper Thompson and Sigman compare the effects of no grouping (*i.e.*, the sample median), grouping with fixed-size intervals, and grouping with data-dependent-sized intervals on medians and associated MHS variance estimates. Their empirical study shows that the data-dependent-sized intervals yielded variance estimates with the smallest bias, the best stability, and the best confidence intervals.

McLaren and Steel consider the implications of different overlap patterns on the sampling variance of seasonally adjusted and trend estimates obtained from time series based on sample surveys by using the Census X-11 and X-11-ARIMA seasonal adjustment methods. They show that the "in for 8", "in for 6", "in for 4, out for 4, in for 4" rotation patterns are sensible if the one month change in seasonally adjusted estimates are the key statistics to be analyzed. If, however, the key statistics are the trend level and the difference between two consecutive trend estimates, then the "in for 1, out for 2, in for 1, for a total of 8 months" is a preferable rotation pattern to reduce the sampling variance. They also show that the "in for 2, out for 2, in for 2, for a total of 8 months is a reasonable compromise if the level and one months change in seasonally adjusted and trend estimates are both considered important.

You and Rao present hierarchical Bayes multi-level models for small area estimation. The models allow random regression parameters that also depend on small area level covariates. The small area mean is estimated by the posterior mean and the posterior variance is taken as a measure of precision. Three variance models are considered: fixed equal, fixed unequal, and random. Details of Gibbs sampling for these models are presented and used for inference. Procedures are illustrated using county level household income data from Brazil.

Okafor and Lee consider a two phase sampling scenario, where a subsample of the non respondents at the second phase are revisited according to a fixed sampling rate. Based on this scheme, modified versions of the ratio and regression estimators are suggested. Optimal values for the sample sizes and the fixed sampling rate aredetermined, based on cost functions, so as to minimize variance. In addition variances and their estimators are given. A small empirical study looks at the relative efficiencies of the modified ratio and regression estimators relative to the standard Hansen-Hurwitz estimator.

Pickery and Loosveldt bring an important analytical technique to the study of item non-response. Their models present a more complete picture of the factors affecting item non-response than in previous work in this area. One important aspect of this approach is that the authors make a separation between interviewer/respondent specific variation, variation attributable to interviewer/ respondent characteristics and error variance.

Fuchs investigates the affect that screen design and question order have on interviewer behavior in a Computer Assisted Interview (CAI) environment. Through the use of experiments under laboratory conditions, it has been shown that screen design and question order do affect interviewer performance. In his paper, Fuchs presents results from a field experiment which tests two different screen designs together with two different question orders in a 2x2 factor design. These results were based on time measures that were built into the CATI application and from 234 randomly selected interviews that were video taped and analyzed according to a coding scheme.

M.P. Singh

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 119-120
Statistics Canada

Leslie Kish – A Life of Giving

IVAN P. FELLEGI[1]

1. INTRODUCTION

I cannot believe that I am writing an article in memory of Leslie Kish. Just a few months ago I wrote a partly humorous little speech on the occasion of his 90th birthday celebration. I jokingly asked why are we making such a fuss about a 90th birthday – after all the Queen mother just celebrated her 100th. I emphasized that *that* was something. He laughed heartily, with the well known "Kish twinkle" in his eye. I was struck once again by the extent to which he remained fun-loving, vibrant, insightful, in fact *young* in all aspects of behaviour – even if somewhat limited in his mobility. He told me about his forthcoming partial knee replacement operation and confided that his doctor told him that he will either undergo this operation, or he will need to use a walker to get around. Of course, a walker was not to be contemplated: he needed to have his full mobility. And mobility, at 90, meant not just getting around at home but traveling around the world several times a year. He died due post-operative complications, having fought for several weeks with his usual indomitable courage.

In my mind the most characteristic feature of his life was his incessant giving. One of his last acts of giving was to inspire his friends and colleagues to establish the Leslie Kish International Fellows Fund to help students from developing countries obtain training in population sampling.

Leslie was born in 1910 in Poprad, then part of the Austro-Hungarian Empire, now in Slovakia. He used to relate how, at various times throughout history, Poprad belonged to five different countries – an appropriate symbol of his life motivated by a love of people from all parts of the world. In 1925 his parents decided to migrate to the U.S.A – together with hundreds of thousands of other Hungarians who left their country. As the great Hungarian poet Attila Jozsef put it: "one and a half million of our people staggered out to America". Soon after their arrival Leslie's father died. The remaining family of mother and four children had to decide whether they will stay in the U.S.A. They did, but that meant that the two oldest children, including Leslie, who was then 16 years old, would have to work in order to help the others.

Leslie continued his schooling in the evening. By 1937 he was within a year of completing his undergraduate studies. But this 27 year old was once again ready to sacrifice himself in order to help the world improve. He interrupted his studies in order help fight the fascists in Spain as a member of the International Brigade. His love of things Spanish, and of people oppressed, stayed with him forever.

At the end of the Spanish Civil War in 1939 he returned to the United States and completed his studies at City College of New York and received a degree in mathematics. He moved to Washington, where he was fortunate to have become a member of pioneering groups, first at the Bureau of the Census and then at the Department of Agriculture.

Again, he interrupted his career to volunteer for service in the war. In 1947 he finally moved to the University of Michigan at Ann Arbor where, together with a small band of enthusiasts helped found the Institute for Social Research. He said later that he never worked as hard as he did in those early years: obtaining his M.A and Ph.D. while working full time but also finding time to teach.

In statistics, he gave us several superb books. These include the pioneering *Survey Sampling* which became not just a bible of the field (*i.e.*, like the original one, a source of lofty inspiration), as well as a day to day tool of practice. In that sense much of the world's statistical system has embedded in it the hundreds of pearls of practical wisdom of *Survey Sampling*. In 1988 (when Leslie was a young 78) came *Statistical Design for Research* which integrated and organized a lifetime's worth of acquired statistical wisdom. In between, before and after came a stream of articles, lectures and talks. He, sometimes working with others, introduced the concepts into our thinking and the words into our language of *design effects*; he was among the first to explore the issue of inference from complex samples and developed the innovation now known as *balanced repeated replication* (actually with Marty Frankel); was among the pioneers of studying *response errors*; became the apostle of *rolling samples and censuses;* pioneered *controlled selection*; formulated the concept of *multipurpose designs*; did some of the early work on *small area estimation*; and so on. But important as these works are, I think just as crucial were some of his other contributions.

He was one of very few people whose early *applied* work made sampling respectable and admired. In addition to having been one of the founders of what became the *Institute for Survey Research* at Ann Arbor, he taught generations of statisticians, both Americans and foreign ones through the legendary Summer Program for Foreign Statisticians. After his formal retirement he continued to do so through lectures in the Summer Program; through

[1] Ivan P. Fellegi, Chief Statistician, Statistics Canada, 26th floor, section A, R.-H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

decades of editing or contributing to one or another of the questions and answers columns of the Survey Statistician; and through numerous lectures and consulting assignments At international meetings I used to "bump into" his past students and current friends. One no longer "bumps into" them, because they have become completely ubiquitous: I wonder how many better known foreign samplers there are who were *not* at some point Leslie's students. And I do not want to forget about two of my favourites among his many contributions. His years of faithful service to Statistics Canada as a founding member of our Advisory Committee on Statistical Methods; and his ASA presidential address of 1977 (published in *JASA* in March 1978) – the best address that any President of ASA gave in my living memory.

For his accomplishments he received world wide recognition. Of his dozens of awards I will just single out a few: he received an honorary doctorate from the University of Bologna on the occasion of its 900th anniversary, the Samuel Wilks Medal which is ASA's highest recognition, the Henry Russell lectureship which is the highest recognition of University of Michigan, the title Honorary Fellow of the ISI which I regard as a kind of Nobel price in statistics, and perhaps the most personally meaningful for him: a slew of the highest possible recognitions from Hungary (honorary doctorate from the largest university in Budapest, honorary membership in the Hungarian Academy of Sciences and the Officer's Cross of the Order of the Merit).

Over and above what he gave us in statistics, he gave us the phenomenon known as "Leslie Kish, a force of nature": the Spanish Civil War fighter, the philosopher of all things statistical, the ever young agitator for human rights, raconteur, avid reader, author of the best annual Christmas letters, loving husband and father, and lifelong friend to hundreds, perhaps thousands.

When I spoke at his 90th birthday celebration, I ended by saying that I was hoping to be present at Leslie's really big anniversary – the one the Queen Mother had just passed. And that was not just a joke: he was so full of life, it was not only quite possible to contemplate him living to be a hundred, but rather it was impossible to think about the opposite. Unfortunately, he did pass away. His final act of giving was to donate his body to medical research. Wouldn't it be fitting if the resulting work gave us some insight into the human wonder that was Leslie Kish?...

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 121-129
Statistics Canada

Population Size and Total Estimation When Sampling From Incomplete List Frames With Heterogeneous Inclusion Probabilities

DAWN E. HAINES, KENNETH H. POLLOCK and SASTRY G. PANTULA[1]

ABSTRACT

Information from list and area sampling frames is combined to obtain efficient estimates of population size and totals. We consider the case where the probabilities of inclusion on the list frames are heterogeneous and are modeled as a function of covariates. We adapt and modify the methodology of Huggins (1989) and Alho (1990) for modeling auxiliary variables in capture-recapture studies using a logistic regression model. We present the results from a simulation study which compares various estimators of frame size and population totals using the logistic regression approach to modeling heterogeneous inclusion probabilities.

KEY WORDS: Logistic regression; List frame; Area frame; Capture-recapture sampling.

1. INTRODUCTION

In this paper, we estimate population size and totals when information from multiple independent sampling frames is available. Population elements are assumed to have varying probabilities of inclusion for different sampling frames. These heterogeneous inclusion probabilities may depend on a covariate. For example, suppose we are interested in estimating the number of hog farms and the total number of hogs in North Carolina. Covariate measurements such as hog farm acreage or number of employees indicate the size of hog farms. Larger farms may have a higher chance of being included on a list frame than smaller farms. In capture-recapture experiments, animals may have unequal capture probabilities. Capture (inclusion) probabilities for animals may vary with respect to age, sex, size, or species.

List frames are physical listings of sampling units in the target population. Items found on a list frame can include, but are not limited to, names, addresses, telephone numbers, social security numbers, or physical descriptions of locations. These and other miscellaneous stratification variables are used to identify persons, animals, businesses, or other establishments. List and area sampling frames are constructed and maintained to obtain estimates of the unknown population size and totals. Since frame imperfections such as omissions, duplications, and inaccurate recordings are inevitable in any large data collection operation (Hansen, Hurwitz and Madow 1953), various solutions for dealing with frame imperfections have been proposed in the literature. One approach, first developed by Hartley (1962, 1974), combines an incomplete list frame with an area frame. Further theoretical extensions are due to Cochran (1965), Lund (1968), Fuller and Burmeister (1972), and Bosecker and Ford (1976). Haines and Pollock (1998a) apply the dual frame method to a bald eagle population while Haines and Pollock (1998b) present a more general, theoretical approach to combining multiple frames. These two papers do not consider the case where the inclusion probabilities are heterogeneous. Fienberg (1992) presents an annotated bibliography of the capture-recapture literature specifically related to the census undercount problem, including Wolter (1986, 1990), and Cowan and Malec (1986).

The National Agricultural Statistics Service (NASS) currently employs a multi-frame approach for its sampling and estimation of numerous agricultural commodities. NASS collects and summarizes data on crop acreage, livestock, grain production and stocks, costs of production, farm expenditures, and other agricultural items. Fecso, Tortora and Vogel (1986) provide a review of sampling frames for the agricultural sector of the United States while Nealon (1984) details the multiple and area frame estimators used by the U.S. Department of Agriculture. Pollock, Turner and Brown (1994) offer a model-based capture-recapture solution for estimating frame size based on information from two incomplete list frames. According to Cochran (1977), it is often difficult to obtain a list that corresponds exactly to the population of interest. Lists routinely collected for some purpose are usually found to be incomplete, partially illegible, or to contain an unknown amount of duplication. Since list frames are typically incomplete, estimates based solely on list frames may underestimate the population size. Supplementing available information with an area frame sample may provide efficient estimates of the population size and totals.

An area frame is a collection of geographical areas defined by identifiable boundaries. Area frames are often used by survey practitioners in order to attain complete coverage of the target population. Populations such as farms are naturally associated with the land units comprising the area frame. For example, in an agricultural survey, the region of

[1] Dawn E. Haines, U.S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock and Sastry G. Pantula, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

interest is divided into a set of geographic land masses called segments. Segments, which are the sampling units, are then selected using stratified multistage designs (Kott and Vogel 1995). Rules which link farms in the population to segments in the area frame are defined. Once the farms, or reporting units, within each sampled segment are identified, they are personally enumerated and the pertinent data collected. Nealon (1984) provides a detailed description of the open, closed, and weighted segment estimators. Faulkenberry and Garoui (1991) formulate additional estimators specifically designed for area frames. More complex construction and sampling methods for area frames are discussed in Fecso *et al.* (1986). Area sampling and subsampling from area frames are considered in detail in Kott and Vogel (1995).

In section 2, we consider independent list frames where the list frame elements have heteroscedastic inclusion probabilities. We discuss methods which provide population size and total estimators when information from list frame(s) and an area frame sample is available. Section 3 summarizes results from a simulation study that compares various estimators of frame (population) size and population totals. Finally, results are summarized and discussed.

2. HETEROSCEDASTIC INCLUSION PROBABILITIES

2.1 Population Size Estimation with List Frames

In capture-recapture experiments, different animals may have different capture probabilities. Similarly, individual elements may have different probabilities of inclusion on a list frame. Different list frames may be viewed as different capture occasions. Model M_h denotes the heterogeneity model in the closed population capture-recapture literature (Otis, Burnham, White and Anderson 1978). In a capture-recapture setting, capture probabilities, though assumed to vary from animal to animal, are assumed to be the same for all trapping occasions. The heterogeneity model may have up to $N + 1$ total parameters, namely N and p_i, $i = 1, ..., N$, where N is the population size and p_i denotes the inclusion probability for the i-th unit. For multiple list frames, this corresponds to the assumption that the inclusion probability p_i for element i is constant over all k list frames, $B_1, B_2, ..., B_k$.

Burnham (1972) and Burnham and Overton (1978, 1979) investigate the problem of estimating N in the capture-recapture setting. The proposed estimator for N given by Burnham (1972) is based on the jackknife method of bias reduction (Quenouille 1956). Chao (1988) develops an alternative moment estimator for this model based on capture frequency data (Pollock 1991). Under certain conditions, Chao's proposed estimator is less biased than Burnham's jackknife estimator. In general, it is difficult to find a completely satisfactory estimator of N under Model M_h. Otis *et al.* (1978), as a result, suggest that one should design the entire study to minimize heterogeneity. Norris and Pollock (1996) propose a nonparametric MLE which is still not totally satisfactory.

In capture-recapture experiments, the model expressed as Model M-th allows inclusion probabilities to vary both by trapping occasion (list frame) and individual. Define p_{ij} as the inclusion probability of the i-th element on the j-th list frame. Model M-th is obviously not easy to estimate since it can have up to $tN + 1$ parameters where $t = k$, the number of list frames. Chao, Lee and Jeng (1992), using the idea of sample coverage, propose a nonparametric method of estimating the population size for Model M-th.

An alternative to the nonparametric approach is to model the inclusion probabilities as a function of an auxiliary variable. Pollock, Hines and Nichols (1984), Huggins (1989), and Alho (1990) address the role of auxiliary variables in capture-recapture experiments with unequal capture (inclusion) probabilities. The closed population capture-recapture experiments have $i = 1, ..., N$ individuals and $j = 1, ..., t$ trapping occasions. Again, the $j = 1, ..., t$ trapping occasions are similar to $t = k$, the number of independent list frames. Huggins (1989) and Alho (1990) propose a conditional estimation procedure for estimating the size of a closed population based on one capture and a single recapture. Both of these papers assume the logistic model for the inclusion probabilities, given by

$$p_{ij} = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \exp(\alpha_j + \beta_j x_i)}, \tag{1}$$

where x_i is a covariate α_j and β_j are unknown parameters. Note that this parameterization yields $0 \le p_{ij} \le 1$ for all values of α_j and β_j. For $\beta_j > 0$, the inclusion probability increases with the covariate. This parameterization is different from the probability proportional to size (pps) sampling where p_{ij} is assumed to be proportional to x_i. The MLEs of α_j and β_j can be obtained using the likelihood conditioned on the unit being on at least one list frame. Haines (1997) derives the conditional likelihood function for three independent list frames.

Treating each individual as a separate stratum, define the following indicator variables for $i = 1, ..., N$:

$$u_{ij} = \begin{cases} 1 & \text{individual } i \text{ belongs to frame } j \text{ only} \\ 0 & \text{otherwise} \end{cases}$$

$$j = B_1, B_2,$$

and

$$a_i = \begin{cases} 1 & \text{individual } i \text{ belongs to both frames} \\ 0 & \text{otherwise.} \end{cases}$$

The value of the expression

$$M_i = u_{iB_1} + u_{iB_2} + a_i \tag{2}$$

is one if individual i is included on at least one of the two frames and zero otherwise.

Alho (1990) presents the conditional likelihood function for two list frames as

$$\frac{\exp\left\{\alpha_{B_1} N_{B_1} + \alpha_{B_2} N_{B_2} + \beta_{B_1} \sum_{i \in B_1} x_i + \beta_{B_2} \sum_{i \in B_2} x_i\right\}}{\Pi_{M_i = 1} K_i(\theta)}, \quad (3)$$

where

$$K_i(\theta) = \exp\{\alpha_{B_1} + \beta_{B_1} x_i\} + \exp\{\alpha_{B_2} + \beta_{B_2} x_i\} + \exp\{\alpha_{B_1} + \alpha_{B_2} + (\beta_{B_1} + \beta_{B_2}) x_i\}$$

and $\theta = (\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}, \beta_{B_2})'$. Alho (1990) uses an iterative procedure based on the sufficient statistics to maximize (3) while we implement Newton's method to calculate conditional MLEs of θ, denoted $\hat{\theta} = (\hat{\alpha}_{B_1}, \hat{\beta}_{B_1}, \hat{\alpha}_{B_2}, \hat{\beta}_{B_2})'$. See Appendix A of Haines (1997) for details on Newton's method. The estimated probability that individual i is included on at least one list frame is denoted

$$\hat{\pi}_i = 1 - \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_1} + \hat{\beta}_{B_1} x_i)}\right) \times \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_2} + \hat{\beta}_{B_2} x_i)}\right) = \pi_i(\hat{\theta}), \quad (4)$$

where

$$\hat{p}_{ij} = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j x_i)}{1 + \exp(\hat{\alpha}_j + \hat{\beta}_j x_i)},$$

$$i = 1, \ldots, N \text{ and } j = B_1, B_2. \quad (5)$$

If θ were known, the Horvitz-Thompson estimator of N is $\hat{N} = \sum_{M_i = 1} 1/\pi_i$ (Horvitz and Thompson 1952). From Cochran (1977), the variance of $\hat{N}$ is

$$V(\hat{N}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i}. \quad (6)$$

An estimate of the variance of $\hat{N}$ is

$$\hat{V}(\hat{N}) = \sum_{M_i = 1} \frac{1 - \pi_i}{\pi_i^2}.$$

Since θ is unknown, we consider the population size estimate given by $\hat{\hat{N}} = \sum_{M_i = 1} 1/\hat{\pi}_i$, where $\hat{\pi}_i$ is defined in (4). An estimate of the variance of $\hat{\hat{N}}$ is derived using Taylor's method and has the form

$$\hat{V}(\hat{\hat{N}}) = \sum_{M_i = 1} \frac{1 - \pi_i(\hat{\theta})}{\pi_i^2(\hat{\theta})} + \hat{A} \hat{\Sigma}(\hat{\theta}) \hat{A}', \quad (7)$$

where

$$\hat{A} = \sum_{M_i = 1} \left[\frac{1}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'}\right]$$

and $\hat{\Sigma}(\hat{\theta})$ is the inverse of the Hessian matrix. The second term in (7) is due to estimating $\pi_i(\theta)$ by $\pi_i(\hat{\theta})$.

Another population size estimator commonly used in capture-recapture experiments is the Lincoln-Petersen estimator. This classic estimator is due to Lincoln (1930) and Petersen (1896) and has the form

$$\hat{N}_{L-P} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}},$$

where N_{B_1} and N_{B_2} denote the size of list frames B_1 and B_2, respectively, and $N_{b_1 b_2}$ denotes the number of units common to both frames. This is a simple method of moments estimator based on the assumption that all units have homogeneous inclusion probabilities for each of the two independent list frames. It is possible for the denominator $N_{b_1 b_2}$ to be zero. Chapman (1951) proposed a modified version of the Lincoln-Petersen estimator, given by

$$\hat{N}_{CH} = \frac{(N_{B_1} + 1)(N_{B_2} + 1)}{(N_{b_1 b_2} + 1)} - 1. \quad (8)$$

This estimator is less biased than the Lincoln-Petersen estimator (Chapman 1951). According to Sekar and Deming (1949), the asymptotic standard error of $\hat{N}_{CH}$ is

$$\sqrt{\hat{V}(\hat{N}_{CH})} = \sqrt{\frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}},$$

where N_{b_1} and N_{b_2} denote the number of units belonging only to list frames B_1 and B_2, respectively.

The Lincoln-Petersen estimator is the unconditional maximum likelihood estimator of the population size N when there are two independent list frames and the inclusion probabilities are homogeneous. Haines (1997) extends the estimation procedures to k list frames, each with homogeneous inclusion probabilities. This estimator, however, is not appropriate when the inclusion probabilities are heterogeneous. See the simulation results in section 3.

2.2 Population Size Estimation with Area and List Frames

Suppose we have access to an area frame in addition to two list frames, B_1 and B_2. The area frame consists of U_A segments that cover the entire population. A simple random sample of u_A segments is selected. We assume that all units in the sampled segments are observed. The probability of inclusion in the area frame sample is the same for all units and is the known quantity $p_A = u_A / U_A$. Next, we maximize the conditional likelihood (3) with respect to θ and calculate the estimated probability that individual i is included on at least one list frame or the area frame. This probability is denoted $\hat{\pi}^*_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i)$. The probabilities $\hat{\pi}_i$ and $\hat{p}_{ij}$ are defined in (4) and (5), respectively. An estimated Horvitz-Thompson estimator for population size is

$$\hat{\hat{N}} = \sum_{i \in \text{sample}} \frac{1}{\tilde{\pi}_i}. \tag{9}$$

This estimator can easily be extended to the case with k list frames, $B_1, \ldots, B_k$, and an independent area frame.

From Cochran (1977), an estimate of the variance of $\hat{\hat{N}}$ is given by

$$\hat{V}(\hat{\hat{N}}) = \sum_{M_i=1} \frac{1-\tilde{\pi}_i}{\tilde{\pi}_i^2} + 2\sum_{i<l}\sum_{l} \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i\tilde{\pi}_l)}{\tilde{\pi}_{il}\tilde{\pi}_i\tilde{\pi}_l}$$

$$+ \hat{A}\hat{\sum}\hat{A}', \tag{10}$$

where $\hat{A}$ is defined in (7) and $\hat{\sum}$ is the inverse of the Hessian matrix. The variance formula for $\hat{N}$ in (6) and its estimate are valid only when π_{il}, the probability that units i and l are included in the sample, is equal to $\pi_i\pi_l$. When an area frame sample is included, however, π_{il} is not necessarily equal to $\pi_i\pi_l$. Suppose units i and l belong to the same area frame segment. In this case, units i and l are both included or not included in the sample, depending on whether their corresponding segment is selected or not. It can be shown that the joint inclusion probability, π_{il}, can be estimated as

$$\tilde{\pi}_{il} = \begin{cases} \tilde{\pi}_i\tilde{\pi}_l & \text{if } i \text{ and } l \text{ belong to different area segments} \\ p_A + \hat{\pi}_i\hat{\pi}_l(1-p_A) & \text{if } i \text{ and } l \text{ belong to the same area segment} \end{cases} \tag{11}$$

where $\hat{\pi}_i$ is defined in (4) and $\tilde{\pi}_i = \hat{\pi}_i + p_A(1-\hat{\pi}_i)$. Hence, when i and l belong to the same segment, $\tilde{\pi}_{il} \neq \tilde{\pi}_i\tilde{\pi}_l$. However, if p_A is small and $\hat{\pi}_i$ and $\hat{\pi}_l$ are large, then $(\tilde{\pi}_{il} - \tilde{\pi}_i\tilde{\pi}_l)$ will be close to zero.

2.3 Population Total Estimation with List Frames

Suppose y_i values are available for all elements on two independent list frames B_1 and B_2. If θ were known, an estimate of the population total, Y, is the Horvitz-Thompson estimator

$$\hat{Y}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\theta)}. \tag{12}$$

According to Cochran (1977), the estimated variance of $\hat{Y}_{H-T}$ is

$$\hat{V}(\hat{Y}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1-\pi_i(\theta))}{\pi_i^2(\theta)}.$$

When θ is unknown and is estimated by $\hat{\theta}$, an estimate for the population total is

$$\hat{\hat{Y}}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\hat{\theta})}.$$

An estimate of the variance of $\hat{\hat{Y}}_{H-T}$ is derived using Taylor's method and has the form

$$\hat{V}(\hat{\hat{Y}}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1-\pi_i(\hat{\theta}))}{\pi_i^2(\hat{\theta})} + \hat{B}\hat{\sum}(\hat{\theta})\hat{B}',$$

where

$$\hat{B} = \sum_{M_i=1} \left[\frac{y_i}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'} \right]$$

and $\hat{\sum}(\hat{\theta})$ is the inverse of the Hessian matrix evaluated at $\hat{\theta}$. These ideas extend easily to incorporate k independent list frames.

In practice, y_i's may not be observed for all units on the list frames. Consider the case where y_i's are available for only a random sample of n_{B_1} and n_{B_2} units from list frames B_1 and B_2, respectively. By construction, the inclusion probabilities, p_{ij}, vary with the individual i and frame j. However, once individuals are included on a list frame, they are subsampled using simple random sampling. As a result, all elements on list frame B_j have equal chance (n_{B_j}/N_{B_j}) of inclusion in the subsample. Note that we are selecting samples from each list frame rather than drawing a single sample from a combined list frame. Since the list frames are assumed to be independent, the estimated probability the i-th individual is included on at least one of the two list frames is

$$\hat{\pi}_i = \hat{p}_{iB_1}\frac{n_{B_1}}{N_{B_1}} + \hat{p}_{iB_2}\frac{n_{B_2}}{N_{B_2}} - \hat{p}_{iB_1}\hat{p}_{iB_2}\frac{n_{B_1}}{N_{B_1}}\frac{n_{B_2}}{N_{B_2}}. \tag{14}$$

An estimated Horvitz-Thompson estimate of Y is obtained by substituting (14) into (12).

Another estimator of the population total, Y, in this case is

$$\hat{Y} = \hat{N}_{CH}\frac{\sum_{M_i=1} y_i}{N_{b_1} + N_{b_2} + N_{b_1b_2}},$$

which is Chapman's estimator multiplied by the mean of the responses for those elements included on at least one list frame subsample. Again, this estimator is valid only when the inclusion probabilities are homogeneous. There are $N_{b_1} + N_{b_2} + N_{b_1b_2}$ unique elements in frames B_1 and B_2. A similar estimator can be defined when information is available only for subsamples from the list frames.

2.4 Population Total Estimation with Area and List Frames

Consider the case where, in addition to y_i values for the units on the list frames (or subsamples from list frames), y_i values are available for all elements in a random sample of segments from an area frame. Inclusion of the area frame information results in the estimated inclusion probability for the i-th individual, namely

$$\tilde{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i), \quad (15)$$

where $\hat{\pi}_i$ is defined in (4) or (14), depending on whether y_i is observed for all units on the list frames or only for a subsample of units, respectively. An estimated Horvitz-Thompson estimator of the population total in this case is

$$\hat{Y}_{H-T} = \sum_{i \in \text{sample}} \frac{y_i}{\tilde{\pi}_i}. \quad (16)$$

An estimate of the variance of $\hat{Y}_{H-T}$ is given by

$$\hat{V}(\hat{Y}_{H-T}) = \sum_{M_i = 1} \frac{y_i^2(1 - \tilde{\pi}_i)}{\tilde{\pi}_i^2} + 2 \sum_{i<l} \sum \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l)}{\tilde{\pi}_{il} \tilde{\pi}_i \tilde{\pi}_l} y_i y_l + \hat{\boldsymbol{B}} \hat{\sum} \hat{\boldsymbol{B}}', \quad (17)$$

where $\tilde{\pi}_{il}$ is defined in (11) and $\hat{\boldsymbol{B}}$ and $\hat{\sum}$ are defined in (13).

3. SIMULATION STUDY

3.1 Assumptions of the Study

To study the properties of population size and total estimators, Haines (1997) considered eighty different models. Details for only two of those models are presented here. One assumption made is that the inclusion probabilities for two list frames depend on a covariate x_i. Secondly, we assume that the covariate may be correlated with the response variable y_i. Also, we assume that x_i and y_i are lognormally distributed with correlation ρ_{xy}. The lognormal distribution is utilized which allows for a skewed distribution of covariates. We generate x_i as e^{u_i} and y_i as e^{v_i}, where u_i and v_i are generated as bivariate normal random variables with zero means, unit variances, and correlation ρ_{uv}. It can be shown that $\rho_{uv} = \log[\rho_{xy}(e-1) + 1]$.

Consider a population of size N. Assume that there are two independent list frames, B_1 and B_2, and an area frame, A. The area frame is assumed to be complete in the sense that it covers the entire population. A sample of area frame segments is selected and the units within each area segment are observed. Let p_A denote the inclusion probability for any element to be included in the area frame sample, where p_A is assumed to be the same for all individuals.

The probability that the i-th element is included on the j-th list frame is given by the logistic regression model (1) for $i = 1, ..., N$ and $j = B_1, B_2$. We assume the probability that the i-th element is included on list frame B_1 is independent of its inclusion status on list frame B_2 and the area frame sample.

3.2 Parameter Settings

We consider various parameter values. For the population size, N, we take $N = 300$ or 1,000. We use $\rho_{xy} = -0.3, 0.0, 0.5$, and 1 corresponding to negative, zero, positive, and perfect correlation between the response variable and the covariate. Here, $\rho_{xy} = 1$ corresponds to $x_i = y_i$, indicating that the inclusion probability is directly related to the response variable.

For each of the above $2 \times 4 = 8$ parameter settings of N and ρ_{xy}, we consider two models corresponding to different choices of $\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}$, and β_{B_2}. Recall that $E(x_i) = E[e^{u_i}] = e^{0.5}$. Consider an element with covariate value given by the mean value $e^{0.5}$. The probability that this element is included on the j-th list frame is

$$p_j^{(E)} = \frac{\exp(\alpha_j + \beta_j e^{0.5})}{1 + \exp(\alpha_j + \beta_j e^{0.5})}, \quad j = B_1, B_2.$$

If $\alpha_j = -\beta_j e^{0.5}$, then this element has a 50% chance of being included on list frame j. We use this relationship in Model 1.

Extending the above idea, if we set

$$\alpha_j = \log\left(\frac{p}{1-p}\right) - \beta_j e^{0.5},$$

then the unit with mean covariate value has probability p of being included on list frame j. If we assume that the inclusion probabilities are the same for list frames B_1 and B_2, then the chance of being included on at least one of the two list frames is given by $1 - (1-p)^2$. This relationship is used in Model 2. Specific values of α_j and β_j for the two models are summarized in Table 1.

Table 1
Summary of Model Parameters

Model	α_{B_1}	β_{B_1}	α_{B_2}	β_{B_2}	$p_{B_1}^{(E)}$	$p_{B_2}^{(E)}$	$1-(1-p_{B_1}^{(E)}).(1-p_{B_2}^{(E)})$
1	0	0	0	0	0.5	0.5	0.75
2	-0.5478	0.8	-0.5478	0.8	0.6838	0.6838	0.90

For each of the $2 \times 4 \times 2 = 16$ models, we consider three p_A values given by 0, 0.05, and 0.20. Here, $p_A = 0$ corresponds to using only the information from list frames B_1 and B_2.

3.3 Generation of the Data

For each of the above sixteen models, we first generate (x_i, y_i) using the bivariate lognormal distribution for $i = 1, ..., N$. We then "generate" (identify) the units that belong to list frames B_1 and B_2. We use the probability p_{ij} to include the i-th element on list frame j. Finally, using $p_A = 0.05$, we identify the elements belonging to area frame A. We repeat the process for the case $p_A = 0.20$. For each parametric combination, we generate 1,000 Monte Carlo replications.

3.4 Estimators

For population size, we consider Chapman's estimator, $\hat{N}_{CH}$, given in (8). This estimator assumes that $\beta_{B_1} = \beta_{B_2} = 0$ and does not utilize the information from the area frame sample. We also consider the estimated Horvitz-Thompson estimators discussed in section 2.

For estimating the population total of a response variable, we consider the case where the response is observed for all list frame elements. Elements in an area frame are sampled with probabilities $p_A = 0, 0.05$, and 0.20. We do not consider population total estimates based on subsamples from each list frame. The population total estimate, $\hat{Y}_{p_A}$, has the same form as (16) with $\tilde{\pi}_i$ defined in (15). Similarly, the population size estimate, $\hat{N}_{p_A}$, has the same form as (9).

The estimator

$$\hat{Y}_{\text{CHs},p_A} = \hat{N}_{\text{CH}}\,\bar{y}_{(p_A)},\ p_A = 0, 0.05, 0.20$$

is also considered where $\bar{y}_{(p_A)}$ is the sample mean of the y_i's included in the "sample." The performance of $\hat{Y}_{\text{CHs},p_A}$ is dependent on $\hat{N}_{\text{CH}}$, which was observed to underestimate N considerably for Model 2. The results for $\hat{Y}_{\text{CHs},p_A}$ are not included here. Another design-unbiased estimator of Y is given by

$$\hat{Y}_A = \sum_{i \in \text{"area sample"}} \frac{y_i}{p_A}.$$

This is the Horvitz-Thompson estimator based on the area frame sample alone. Since complete enumeration of area segments is expensive, p_A is typically small in practice. For small p_A, $\hat{Y}_A$ is expected to have a much larger variance than $\hat{Y}_{p_A}$ since the estimator $\hat{Y}_{p_A}$ includes information from list frames in addition to information from the area frame samples. Hence, results for $\hat{Y}_A$ are not included.

3.5 Estimated Variance of the Estimator

In our simulation study the values of p_A considered are very small. In contrast, the probability of inclusion on at least one of the list frames is large for each individual. As a result, $\tilde{\pi}_i$ is close to $\hat{\pi}_i$ and $\tilde{\pi}_{il}$ in (11) is close to $\tilde{\pi}_i \tilde{\pi}_l$. Hence, the second term in equations (10) and (17), involving $\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l$, are expected to be small. We have not included this term in our estimate of the variance. Despite this omission, we observe that the estimated variance is very close to the empirical variance of the estimator for the models we consider.

3.6 Summary Statistics

For the population size estimates, we present results averaged over the 4,000 replications corresponding to the four values of ρ_{xy} and 1,000 Monte Carlo replications for each ρ_{xy}. For each model, we summarize the mean and standard deviation of the estimates, average of the estimated standard errors of the estimators, the percent relative root mean square error (% RRMSE), and the empirical coverage probabilities of a 95% confidence interval. These measures are all standardized by the population size N. We report results for Models 1 and 2 in Tables 2 and 3, respectively.

Table 2
Population Size Estimates for Model 1

$N = 300$	$\hat{N}_{CH}$	$\hat{N}_0$	$\hat{N}_{0.05}$	$\hat{N}_{0.20}$
Average of estimates divided by N	0.999	1.011	1.007	1.004
Standard deviation of estimates divided by N	0.059	0.077	0.059	0.048
Average of estimated standard deviation of estimator divided by N	0.059	0.072	0.059	0.047
% RRMSE	0.003	0.006	0.004	0.002
Coverage	0.947	0.955	0.957	0.950
$N = 1{,}000$				
Average of estimates divided by N	1.000	1.003	1.002	1.002
Standard deviation of estimates divided by N	0.031	0.035	0.030	0.025
Average of estimated standard deviation of estimator divided by N	0.032	0.034	0.030	0.025
% RRMSE	0.001	0.001	0.001	0.001
Coverage	0.954	0.959	0.958	0.956

Table 3
Population Size Estimates for Model 2

$N = 300$	$\hat{N}_{CH}$	$\hat{N}_0$	$\hat{N}_{0.05}$	$\hat{N}_{0.20}$
Average of estimates divided by N	0.922	1.006	1.005	1.003
Standard deviation of estimates divided by N	0.032	0.052	0.049	0.040
Average of estimated standard deviation of estimator divided by N	0.028	0.052	0.048	0.040
% RRMSE	0.007	0.003	0.002	0.002
Coverage	0.271	0.953	0.954	0.951
$N = 1{,}000$				
Average of estimates divided by N	0.921	1.001	1.001	1.001
Standard deviation of estimates divided by N	0.018	0.028	0.027	0.022
Average of estimated standard deviation of estimator divided by N	0.015	0.027	0.026	0.021
% RRMSE	0.007	0.0008	0.0007	0.0005
Coverage	0.009	0.949	0.949	0.949

Similarly, for the population total estimates, we present summary statistics averaged over the 1,000 replications corresponding to each parametric combination. We summarize the mean and standard deviation of the estimates as well as the average of the estimated standard errors of the estimators, where the estimates are scaled by the true total (Y) for that replicate. In other words, for each replicate we divide the estimate by its replicate total, Y. We then compute the mean and the standard deviations of these standardized estimates. Similarly, for each replicate, we compute the estimated standard error of the total estimator divided by the total for the replicate and then compute the average of these standardized values. We report these because the totals change from replicate to replicate.

Finally, we report the coverage probabilities of the 95% confidence intervals for the total. Results for Models 1 and 2 are respectively presented in Tables 4 and 5.

3.7 Conclusions

3.7.1 Population Size Estimation

In Model 1, the inclusion probabilities do not depend on the covariate. In this case, Chapman's estimator $\hat{N}_{CH}$ is very close to the maximum likelihood (Lincoln-Petersen) estimator and hence is expected to perform better than $\hat{N}_0$. The estimator $\hat{N}_0$ loses efficiency since it estimates the parameters α_{B_1}, β_{B_1}, α_{B_2}, and β_{B_2}, which have the value zero in this model. The estimator $\hat{N}_{0.05}$ has about the same efficiency as $\hat{N}_{CH}$. The bias in all the estimates is minimal. For Model 1, we notice that the average of the estimated standard deviation is close to the standard deviation of the estimates. This indicates that the standard error estimate we use performs well. Also, we notice that the empirical coverage probabilities are all within three standard errors of 0.95. That is, all of the empirical coverage probabilities are within $(0.95 \pm 3\,[0.95 \times 0.05/4{,}000]^{1/2}) = (0.94, 0.96)$.

Table 4
Population Total Subsampling Estimates Scaled by Y for Model 1

		$N = 300$			$N = 1{,}000$		
ρ_{xy}		$\hat{Y}_0/Y$	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$	$\hat{Y}_0/Y$	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$
-0.3	Average of estimates	1.004	1.003	1.001	1.002	1.002	1.001
	Standard deviation of estimates	0.077	0.073	0.062	0.042	0.040	0.035
	Average of estimated standard error	0.076	0.072	0.061	0.041	0.039	0.033
	Coverage	0.953	0.951	0.949	0.946	0.942	0.942
0	Average of estimates	1.013	1.012	1.008	1.001	1.001	1.001
	Standard deviation of estimates	0.080	0.070	0.059	0.041	0.039	0.033
	Average of estimated standard error	0.081	0.072	0.060	0.040	0.038	0.033
	Coverage	0.951	0.954	0.951	0.944	0.942	0.946
0.5	Average of estimates	1.053	1.018	1.009	1.004	1.003	1.002
	Standard deviation of estimates	0.586	0.104	0.072	0.057	0.045	0.037
	Average of estimated standard error	0.233	0.094	0.070	0.051	0.045	0.036
	Coverage	0.950	0.951	0.945	0.950	0.955	0.955
1.0	Average of estimates	1.064	1.030	1.013	1.013	1.009	1.006
	Standard deviation of estimates	0.515	0.162	0.090	0.094	0.066	0.047
	Average of estimated standard error	0.277	0.128	0.086	0.070	0.059	0.046
	Coverage	0.930	0.929	0.930	0.946	0.949	0.951

Table 5
Population Total Subsampling Estimates Scaled by Y for Model 2

		$N = 300$			$N = 1{,}000$		
ρ_{xy}		$\hat{Y}_0/Y$	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$	$\hat{Y}_0/Y$	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$
-0.3	Average of estimates	1.010	1.009	1.006	1.003	1.003	1.002
	Standard deviation of estimates	0.098	0.092	0.078	0.052	0.049	0.041
	Average of estimated standard error	0.094	0.089	0.074	0.051	0.048	0.040
	Coverage	0.935	0.926	0.931	0.952	0.955	0.955
0	Average of estimates	1.008	1.007	1.005	1.002	1.002	1.001
	Standard deviation of estimates	0.065	0.062	0.050	0.034	0.032	0.027
	Average of estimated standard error	0.064	0.061	0.051	0.034	0.032	0.028
	Coverage	0.953	0.950	0.952	0.947	0.951	0.955
0.5	Average of estimates	1.002	1.002	1.001	1.001	1.001	1.001
	Standard deviation of estimates	0.035	0.033	0.028	0.019	0.018	0.015
	Average of estimated standard error	0.035	0.034	0.029	0.019	0.018	0.016
	Coverage	0.954	0.950	0.951	0.965	0.967	0.965
1.0	Average of estimates	1.001	1.001	1.001	1.000	1.000	1.000
	Standard deviation of estimates	0.021	0.020	0.017	0.012	0.011	0.010
	Average of estimated standard error	0.021	0.020	0.017	0.012	0.011	0.009
	Coverage	0.952	0.949	0.954	0.947	0.947	0.943

For Model 2, the inclusion probability is a function of the covariate. As a result, $\hat{N}_{CH}$ is not an appropriate estimator for N. We observe that $\hat{N}_{CH}$ significantly underestimates the true population size. On the other hand, $\hat{N}_{p_A}$ provides a good estimate of N. The bias in $\hat{N}_{p_A}$ decreases as p_A increases in Model 2. Further, the relative bias decreases as the population size increases.

As expected, the standard deviation of $\hat{N}_{p_A}$ decreases as the area frame inclusion probability p_A increases. For example, in Model 1 where $N = 300$, the inclusion of a 5% area frame sample reduces the relative standard deviation from 0.077 to 0.059, a 23% reduction. When a 20% area frame sample is utilized, the relative standard deviation decreases from 0.077 to 0.048, a 38% reduction. When $N = 1{,}000$, the inclusion of a 5% area frame sample decreases the relative standard deviation from 0.035 to 0.030, a 14% reduction. Increasing the area frame sample to 20% reduces the relative standard deviation from 0.035 to 0.025, a decrease of 29%. Generally speaking, the relative standard errors decrease as population size increases. Although the average of the estimated standard error of $\hat{N}_{p_A}$ is smaller than the empirical standard deviation, the difference is relatively small. Also, the coverage probabilities of the 95% confidence interval based on $\hat{N}_{p_A}$ are very close to 0.95. In contrast, the coverage probabilities of the 95% confidence interval based on $\hat{N}_{CH}$ are 0.271 and 0.009 for $N = 300$ and $N = 1{,}000$, respectively.

Based on our simulations, we recommend the use of $\hat{N}_{p_A}$ with a large value of p_A. The choice of p_A is determined in practice by area frame sampling costs, which are not taken into consideration in our study.

3.7.2 Population Total Estimation

For population totals, we observe results that are very similar to what we observed for the population size. In general, relative biases and standard errors decrease as p_A increases and as the population size increases. We also notice that the average relative estimated standard error is very close to the empirical standard deviation of the standardized estimator standardized by the total. This suggests that the approximate standard error formula in (7) is a good estimate of the standard error. Note also that the empirical coverage probabilities are mostly within three standard errors of 0.95. That is, most of the empirical coverage probabilities fall within $(0.95 \pm 3[0.95 \times 0.05/1{,}000]^{1/2}) = (0.929, 0.971)$.

4. SUMMARY

In this paper, we studied the performance of the estimated Horvitz-Thompson estimator of the population size and total based on samples from area and list frames. We presented methods for estimating the parameters of the logistic regression model for the inclusion probabilities. Though numerous models and other estimators are considered in Haines (1997), we presented simulation study results for only two models and a few estimators.

We believe the methods used in this paper are potentially very useful to survey researchers because list frame incompleteness is a fact of life. Our results are among the first to suggest a method of estimating population totals which account for incompleteness and model the inclusion probabilities as a function of the covariates.

ACKNOWLEDGEMENTS

The authors thank the editor and an associate editor for their comments which improved the content and presentation of this paper. We also wish to thank Christine Bunck, BEST Program Manager, Biological Resources Division, U.S. Geological Survey, for financial support of this research through a research work order to North Carolina State University. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

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

BURNHAM, K.P. (1972). Estimation of Population Size in Multiple Capture Studies when Capture Probabilities Vary Among Animals. Ph. D. thesis, Oregon State University.

BURNHAM, K.P., and OVERTON, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65, 625-633.

BURNHAM, K.P., and OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.

CHAO, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, 52, 295-300.

CHAO, A., LEE, S.-M., and JENG, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48, 201-216.

CHAPMAN, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. University of California, University of California Publication in Statistics.

COCHRAN, R.S. (1965). Theory and Applications of Multiple Frame Surveys. Ph.D. thesis, Iowa State University.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd edition. New York: John Wiley & Sons.

COWAN, C.D., and MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

FAULKENBERRY, G.D., and GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.

FECSO, R., TORTORA, R.D., and VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.

FIENBERG, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18, 143-154.

FULLER, W.A., and BURMEISTER, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Ph.D. thesis, North Carolina State University.

HAINES, D.E., and POLLOCK, K.H. (1998a). Combining multiple frames to estimate population size and totals. *Survey Methodology*, 24, 79-88.

HAINES, D.E., and POLLOCK, K.H. (1998b). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics*, 5, 245-256.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory 1*. New York: John Wiley & Sons.

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

HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 3, C, 99-118.

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

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

KOTT, P.S., and VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed. B.G. Cox.). New York: John Wiley & Sons. 185-203.

LINCOLN, F.C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. U.S. Department of Agriculture, Circular, 118.

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

NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators, Staff Report 80. U.S. Department of Agriculture, Statistical Reporting Service, Washington, D. C.

NORRIS, J.L., and POLLOCK, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52, 639-649.

OTIS, D.L., BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.

PETERSEN, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea, *Rep. Danish Biol. Sta.*, 6, 1-48.

POLLOCK, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Association*, 86, 225-238.

POLLOCK, K.H., HINES, J.E., and NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.

POLLOCK, K.H., TURNER, S.C., and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117-124.

QUENOUILLE, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 131-136
Statistics Canada

An Estimation Method for Nonignorable Nonresponse

JEAN-FRANÇOIS BEAUMONT[1]

ABSTRACT

When a survey response mechanism depends on a variable of interest measured within the same survey and observed for only part of the sample, the situation is one of nongnorable nonresponse. In such a situation, ignoring the nonresponse can generate significant bias in the estimation of a mean or of a total. To solve this problem, one option is the joint modelling of the response mechanism and the variable of interest, followed by estimation using the maximum likelihood method. The main criticism levelled at this method is that estimation using the maximum likelihood method is based on the hypothesis of error normality for the model involving the variable of interest, and this hypothesis is difficult to verify. In this paper, the author proposes an estimation method that is robust to the hypothesis of normality, so constructed that there is no need to specify the distribution of errors. The method is evaluated using Monte Carlo simulations. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

KEY WORDS: Nonignorable nonresponse; Maximum likelihood; Estimation equations; Regression imputation; Reweighting.

1. INTRODUCTION

When a survey response mechanism depends on a variable of interest measured in the same survey and observed for only part of the sample, the situation is one of nonignorable nonresponse. In measuring income, for example, it may be realistic to assume that low income earners will exhibit a lower tendency to respond than high income earners, or vice versa. Readers will find in Little (1982) a formal definition of the concept of nonignorable nonresponse. In such a situation, ignoring the nonresponse can generate significant bias in the estimation of a mean or of a total. To solve this problem, one option is the joint modelling of the response mechanism and the variable of interest, followed by estimation using the method of maximum likelihood, used for example in Greenlees, Reece and Zieschang (1982), and imputation of the missing values. The main criticism levelled at this method is that estimation using the method of maximum likelihood is based on the hypothesis of error normality for the model involving the variable of interest, and this hypothesis is difficult to verify.

Rancourt, Lee and Särndal (1994) described simple correction factors aimed at reducing the bias generated by nonresponse that is not ignorable without reference to the hypothesis of normality and in the absence of a response mechanism model. These correction factors, however, are only available for ratio imputation.

In this paper, the author proposes an estimation method that is robust with respect to the hypothesis of normality, so constructed that there is no need to specify the distribution of errors. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

In section 2, the problem is defined and some notation is introduced. In section 3, various estimators of the mean of a population are introduced for a variety of hypotheses concerning the response mechanism and the distribution of data. In section 4, an estimation method is proposed for nonignorable nonresponse. In section 5, the author describes the results of a simulation study used to compare the estimators described in the two preceding sections. Finally, the last section contains a brief discussion.

2. NOTATION

In the following, we attempt to estimate the mean of a variable Y for a certain population P. To do so, we select a sample S, and the variable Y is observed for only part of the sample. The sample of respondents is denoted R, and the sample of nonrespondents is denoted O. We assume that there is at least one variable that is observed for all the sampling units and correlated with Y.

The estimator of the mean, $\mu = \sum_{i \in P} Y_i / N$, where N is the size of the population, can be obtained by weighting the respondent units:

$$\mu_P^* = \frac{\sum_{i \in R} w_i w_{R,i}^* Y_i}{\sum_{i \in R} w_i w_{R,i}^*}, \qquad (2.1)$$

where w_i denotes the sampling weights that correspond to the inverse selection probability and $w_{R,i}^*$ denotes the weights that correspond to the estimated inverse response probability. Another estimator of the mean can be obtained by imputing the missing values:

[1] Jean-François Beaumont, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

$$\mu_I^* = \frac{\sum_{i \in R} w_i Y_i + \sum_{i \in O} w_i Y_i^*}{\sum_{i \in S} w_i}, \qquad (2.2)$$

where Y_i^* denotes values that are imputed for the nonrespondent units.

For the sake of simplicity, we assume, in the following, that the sampling weights are constant for all units of the population. Thus, we can eliminate w_i from equations (2.1) and (2.2). We also assume that there is only one observed variable for all sampling units. This variable is denoted X.

3. CURRENT ESTIMATION METHODS

In this section, equations (2.1) and (2.2) are developed under a variety of hypotheses concerning the response mechanism and the distribution of data, and appropriate estimation methods are described. In section (3.1), we assume a uniform response mechanism; in section (3.2), we assume a response mechanism that depends on X, while in section (3.3), we assume a response mechanism that depends on Y. The response mechanisms in sections (3.1) and (3.2) are ignorable, whereas the one in section (3.3) is not ignorable.

3.1 Uniform Response Mechanism

Assuming a uniform response mechanism, we have the same response probability for all sampling units. Thus, estimator (2.1) becomes:

$$\mu_{P,U}^* = \frac{\sum_{i \in R} Y_i}{n_R}, \qquad (3.1)$$

where n_R is the total number of respondents. This estimator is the very same one we would have obtained by using equation (2.2) and by imputing the respondent mean for all nonrespondents.

3.2 Response Mechanism Dependent on X

When the response mechanism depends on variable X (correlated with Y), estimator (3.1) might be strongly biased. It is then preferable to use this variable as additional information for the estimation of mean μ.

Estimator (2.1) can be obtained by replacing $1/w_{R,i}^*$ by the estimated response probability using a logistic regression. A response probability model is therefore needed. If we only have one observed variable (X) for all sampling units, the model can be written as follows:

$$P(R_i = 1 \mid X_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 X_i)]},$$

where α_0 and α_1 are parameters to be estimated (using the maximum likelihood method, for example) and R_i is a dichotomous variable equal to 1 if unit i responds and to 0 otherwise. The estimator of the mean obtained in this way is denoted $\mu_{P,X}^*$.

If we prefer to use estimator (2.2) instead, the missing values can be imputed using the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad (3.2)$$

where β_0 and β_1 are unknown parameters and ε_i is a random error term of zero mean that is not correlated with X_i. The imputed values are given by: $Y_i^* = B_0^* + B_1^* X_i$, where B_0^* and B_1^* are estimates (obtained by means of the method of least squares using the respondent units) of B_0 and B_1 which are in turn estimates of β_0 and β_1. In fact, B_0 and B_1 are the estimates that would have been obtained (using the method of least squares) if we had observed all the units of sample S. The estimator obtained in this way is denoted $\mu_{I,X}^*$. Note that all the models considered in this document are assumed to be valid for all the units of sample S.

We could also add a residual to the imputed values in order to better estimate the variance due to sampling (see for example Gagnon, Lee, Rancourt and Särndal 1996). However, this technique still does not make it possible to estimate the variance due to imputation. Moreover, it tends to produce estimates of the mean that are less precise than if no residual had been added. Since this paper does not deal with variance estimation, we have chosen not to add residuals to the imputed values. This has the added advantage of simplifying the calculation of $\mu_{I,X}^*$.

3.3 Response Mechanism Dependent on Y

All the estimators of the mean discussed so far can be strongly biased when the response mechanism depends on Y (nonignorable response mechanism). For such a response mechanism, the response probability can be modelled as follows:

$$P(R_i = 1 \mid Y_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 Y_i)]}. \qquad (3.3)$$

Since variable Y is only observed for respondent units, it is impossible to obtain an estimate for α_0 and α_1 using the maximum likelihood method. Model (3.2) can also be used. However, the parameter estimates will not be consistent since $E(\varepsilon_i \mid R_i = 1)$ and $E(\varepsilon_i X_i \mid R_i = 1)$ are not zero. Even if we had consistent estimates, the missing values could not be imputed as described in section (3.2) since $E(Y_i \mid R_i = 0, X_i) \neq \beta_0 + \beta_1 X_i$ (Greenlees, Reece and Zieschang 1982). If, for example, the response probability correlates positively with the variable of interest Y, then, for a given value of X, the mean of nonrespondent units will be lower than that of respondent units, and will therefore be lower than the mean of all units taken together. A similar argument can be presented if the response probability correlates negatively with the variable of interest. In fact, it can be shown that

$$E(Y_i \mid R_i = 0, X_i) = \beta_0 + \beta_1 X_i - \frac{\mathrm{cov}(Y_i, p(Y_i) \mid X_i)}{1 - E(p(Y_i) \mid X_i)},$$

where $p(Y_i) = P(R_i = 1 \mid Y_i)$.

The two approaches in section (3.2) are therefore invalid when the response mechanism is not ignorable. In such a situation, a better approach would be to estimate the parameters of models (3.2) and (3.3) simultaneously. The method of maximum likelihood can be used to this end. This method, however, requires as an additional hypothesis that errors ε_i follow a normal distribution (or any other distribution relevant to the type of data analyzed) with constant variance σ^2, and that they be mutually independent. The natural logarithm of the likelihood function l can be written as follows:

$$l = \sum_{i \in R} \ln[p(Y_i) f(Y_i \mid X_i)] + \sum_{i \in O} \ln[1 - E(p(Y_i) \mid X_i)], \quad (3.4)$$

where $f(Y_i \mid X_i)$ is the probability density function of a normal distribution with a mean $\beta_0 + \beta_1 X_i$ and variance σ^2. The method of maximum likelihood consists in finding the parameter values which maximize l. To carry out the maximization, it must be possible to approximate $E(p(Y_i) \mid X_i)$. This can be achieved by using a numerical integration method similar to that of Greenlees, Reece and Zieschang (1982). In this paper, the following approximation (Zeger, Liang and Albert 1988) has been used instead:

$$E(p(Y_i) \mid X_i) \approx \frac{1}{1 + \exp\{-k[\alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i)]\}}, \quad (3.5)$$

where $k = 1/\sqrt{c^2\sigma^2\alpha_1^2 + 1}$ and $c = 16\sqrt{3}/15\pi$. This approximation is based on the hypothesis that errors follow a normal distribution with constant variance. This approximation was preferred to a method of numerical integration because it is simpler and computationnally faster, an advantage that must be considered seriously before any simulation study is undertaken. Finally, equation (3.4) was maximized using the Newton-Raphson algorithm and the NLIN procedure of the SAS software (SAS Institute Inc. 1990).

Once the parameters of models (3.2) and (3.3) have been estimated, estimators of the mean (2.1) or (2.2) can be chosen. Estimator (2.1) is obtained by replacing $w_{R,i}^*$ by $1/p^*(Y_i)$, where $p^*(Y_i)$ is the estimated response probability. This estimator is denoted $\mu_{P,Y,ML}^*$. Estimator (2.2) can be obtained by determining imputed values Y_i^* in such a way that $\sum_{i \in S} e_i^2$ is minimized and that the constraints $\sum_{i \in S} e_i = 0$ and $\sum_{i \in S} e_i X_i = 0$ are met, where $e_i = Y_i - \beta_0^* - \beta_1^* X_i$, for $i \in R$, $e_i = Y_i^* - \beta_0^* - \beta_1^* X_i$, for $i \in O$, and β_0^* and β_1^* are the estimates of β_0 and β_1 respectively. The estimator of the mean can then be written as follows: $\mu_{I,Y,ML}^* = \beta_0^* + \beta_1^* \sum_{i \in S} X_i / n$, where n is the size of sample S. The reasoning behind this approach is that the two previous constraints would have been met if variable Y had been observed for all units in the sample and if this variable had been modelled using model (3.2).

4. PROPOSED METHOD OF ESTIMATION

This section describes the proposed method of estimation for a nonignorable response mechanism (section 4.1), as well as a graphic method (section 4.2) that can be used to verify the error normality hypothesis of model (3.2).

4.1 Method of Estimation for a Response Mechanism Dependent on *Y*

The method of maximum likelihood is valid when errors exhibit a normal distribution and have the same variance. When this hypothesis does not hold, it is preferable to use a more robust method of estimation.

If response probabilities $p(Y_i)$ were known and greater than zero for all sampling units, a robust method of estimation (in terms of both the error normality hypothesis and model 3.2) would consist in minimizing the error sum of squares weighted by the inverse response probability $p(Y_i)$. This minimization is equivalent to solving the system of equations

$$\sum_{i \in R} \frac{1}{p(Y_i)} (Y_i - \beta_0 - \beta_1 X_i) Z_{ik} = 0, \quad k = 1, 2, \quad (4.1)$$

where $Z_{i1} = 1$ and $Z_{i2} = X_i$. This approach is considered robust with respect to the normality hypothesis since the method of least squares does not require that the distribution of errors be specified. Weighting by means of the inverse response probability also provides a certain robustness in terms of model (3.2). In fact, estimators B_0^* and B_1^* obtained using equation (4.1) are consistent with respect to the response mechanism for B_0 and B_1 (which are the estimators of β_0 and β_1 that would have been obtained if there had been no nonresponse) regardless of the validity of the model. A similar argument may be found in Särndal, Swensson and Wretman (1992, p. 519), but in terms of the sample selection mechanism instead of the response mechanism.

Likewise, if the probability density function $f(Y_i \mid X_i)$ was known (not necessarily normal and yet not dependent on the parameters of model 3.3), we could then estimate parameters α_0 and α_1 of model (3.3) using the maximum likelihood method, for example, and solve the system of equations

$$\sum_{i \in R} \frac{\partial}{\partial \alpha_k} \ln[p(Y_i)] + \sum_{i \in O} \frac{\partial}{\partial \alpha_k} \ln[1 - E(p(Y_i) \mid X_i)] = 0, \quad (4.2)$$

for $k = 0$ and $k = 1$.

Thus, the estimates of parameters β_0, β_1, α_0 and α_1 are obtained by solving the unbiased estimation equations (4.1) and (4.2). An algorithm that can be used to find the solution consists in solving alternately the systems of equations (4.1) and (4.2) until convergence is achieved. This requires the possibility of calculating $E(p(Y_i)|X_i)$ in equation (4.2). However, this last expectation requires that the distribution of errors ε_i be known, and in all likelihood it is unknown. To get around this problem, we must use an approximation, and a number of them can be considered, including approximation (3.5). Another option would be to develop a strategy based on the bootstrap method by selecting the respondent units proportionally to their inverse response probability. However, this method requires considerable computer processing time, and is not considered in this paper. Instead, we have chosen the following approximation, obtained by linearizing $p(Y_i)$ using a Taylor series assessed at point $E(Y_i|X_i)$ and by taking the expectation of the first two terms in this series:

$$E(p(Y_i)|X_i) \approx p(E(Y_i|X_i)) = p(\beta_0 + \beta_1 X_i). \quad (4.3)$$

It should be noted that the expectation of the second term in the series is zero. This approximation offers the advantage of requiring only the first moment of the distribution of Y_i conditional on X_i. In this sense, it should be robust with respect to the error normality hypothesis since it does not require that the error distribution be specified. Of course, if the distribution of errors is known or can be properly estimated, it will be possible to find better approximations than (4.3) although, in this case, it may be preferable to use the maximum likelihood method.

Another interesting property of approximation (4.3) is that alternately solving the systems of equations (4.1) and (4.2) might be achieved using the following algorithm:

1. determine initial values for the response probabilities (or for parameters α_0 and α_1), *e.g.*, let $p(Y_i)^{(0)} = 1$ for all the respondent units;
2. let $j = 1$, where j is the number of iterations;
3. solve the system of equations (4.1) by means of the current response probability estimates, $p(Y_i)^{(j-1)}$, using a weighted regression procedure to obtain $\beta_0^{(j)}$ and $\beta_1^{(j)}$;
4. impute the missing values using $Y_i^{(j)} = \beta_0^{(j)} + \beta_1^{(j)} X_i$ for $i \in O$;
5. solve the system of equations (4.2) by using a logistic regression procedure to obtain $p(Y_i)^{(j)}$;
6. stop once convergence has been achieved, otherwise let $j = j + 1$ and return to step 3.

It is sufficient then to simply have a linear regression procedure and a logistic regression procedure to obtain the desired estimates. This algorithm is a very efficient means of finding the solution although, in certain cases, many iterations might be needed before convergence is achieved. In actual practice, it did converge in all cases where it was used. It should also be noted that this algorithm shows certain similarities with the EM algorithm used by Dempster, Laird and Rubin (1977), except that here we do not maximize a likelihood function.

For the simulations in the next section, we selected instead the Newton-Raphson algorithm which converges more rapidly. However, the above-mentioned algorithm had to be used for the few cases in which the Newton-Raphson algorithm met with convergence problems.

The proposed algorithm might be very useful as a means of providing initial values for a more rapid algorithm such as the Newton-Raphson one. The proposed algorithm could simply be used with a not very demanding convergence criterion so that, after only a few iterations, it could provide sufficiently good initial values to ensure convergence of the Newton-Raphson algorithm. In a different context, Beaumont and Demnati (1998) used a similar approach by beginning the iterative process using an algorithm of the EM type so as to provide the initial values for a more rapid algorithm of the Newton-Raphson type. They were able to show empirically that the combination of the two algorithms represents a sound compromise between processing time and efficiency in finding a solution.

As in section (3.3), once the parameters of models (3.2) and (3.3) are estimated, we can select estimators of the mean (2.1) or (2.2). Estimator (2.1) is obtained by replacing $w_{R,i}^*$ by $1/p^*(Y_i)$, where $p^*(Y_i)$ is the estimated response probability. This estimator is denoted $\mu_{P,Y,ROB}^*$. Estimator (2.2) is also obtained as in section (3.3) by determining the imputed values Y_i^* in such a way that $\sum_{i \in S} e_i^2$ is minimized and the constraints $\sum_{i \in S} e_i = 0$ and $\sum_{i \in S} e_i X_i = 0$ are met, where $e_i = Y_i - B_0^* - B_1^* X_i$, for $i \in R$, and $e_i = Y_i^* - B_0^* - B_1^* X_i$, for $i \in O$. This estimator is denoted $\mu_{I,Y,ROB}^*$. The quality of these two estimators of the mean will depend largely on the validity of models (3.2) and (3.3) and on the quality of approximation (4.3).

A modification of step (5) for the algorithm presented in this section was proposed by Beaumont (1999). The results of a simulation study show that this modification provides results that are slightly better than those obtained using the method proposed in this paper. However, this no longer involves using the maximum likelihood method to estimate the parameters of model (3.3), given that $f(Y_i|X_i)$ is known and a logistic regression procedure can no longer be used for step (5). It should nevertheless be mentioned that it is not absolutely necessary to use the method of maximum likelihood to estimate α_0 and α_1, although it is the method preferred in this paper.

4.2 Verifying the Error Normality Hypothesis

In order to use the method of maximum likelihood, we might be interested in verifying the error normality hypothesis (or rather the residual normality hypothesis since the errors are not observed). In the absence of nonresponse, a traditional method (D'Agostino 1986, p. 25, equation 2.11)

consists in producing the graph of $\Phi^{-1}[F_n(e_i)]$ in terms of residuals e_i, for $i \in S$, where $\Phi(.)$ is the distribution function for a random variable having the standard normal distribution, and $F_n(.)$ is the empirical distribution function. Whenever errors exhibit normal distribution, the points in this graph should more or less fall along a line having a slope $1/\sigma$ passing through the origin.

If there is nonresponse, the same strategy can be used as in the previous paragraph, but the empirical distribution function must be estimated using the respondent units. Since the units in the sample respond with unequal probabilities, the estimated empirical distribution function can be given by the formula (Särndal, Swensson and Wretman 1992, p. 199):

$$F_n^*(e_i) = \frac{\sum_{j:j\in R \text{ et } e_j \le e_i} 1/p^*(Y_j)}{\sum_{j\in R} 1/p^*(Y_j)}.$$

Note that, in this last equation, the response probabilities are estimated as opposed to the Särndal, Swensson and Wretman formula, in which selection probabilities are known. Thus, the error normality hypothesis can be verified by producing the graph of $\Phi^{-1}[F_n^*(e_i)]$ in terms of residuals e_i, for $i \in R$. This method will be valid provided that $F_n^*(e_i)$ can correctly estimate $F_n(e_i)$, as is the case when the response probabilities are correctly estimated. When the nonresponse is not ignorable, and when the method of estimation proposed in this paper is used, the response probabilities should be properly estimated if models (3.2) and (3.3) are appropriate along with approximation (4.3).

5. SIMULATION STUDY

In order to compare the estimators of the mean presented in the two previous sections, we carried out a simulation study. We simulated 4 populations with a size of 1,000 according to model (3.2) with $\beta_0 = 2$ and $\beta_1 = 3$. Random variables X_i are independent of one another and they follow an exponential distribution of mean 1. Errors ε_i are independent of one another, are not correlated with the X_i and have a mean of zero and a variance σ^2. In two populations, the errors follow a normal distribution ($\varepsilon_i \sim \text{Nor}(0, \sigma^2)$), and in the other two populations, the errors follow an exponential distribution of mean σ recentred at $0 (\varepsilon_i \sim \text{Exp}(\sigma) - \sigma)$. For each of these distributions, one population has a standard deviation σ equal to 1.5 corresponding to a squared coefficient of correlation (between X and Y) of 80% ($R^2 = 80\%$), and the other has a standard deviation equal to 3 corresponding to a square coefficient of correlation of 50% ($R^2 = 50\%$).

For each population, we simulated 1,000 samples of respondents according to model (3.3) with $\alpha_1 = 0.5$. Parameter α_0 was determined separately for each of the 4 populations, so that the mean response rate would be 70%. This parameter varied between -1.185 and -0.958. Note that we have here a census ($n = N = 1\,000$). The advantage of this is that we can concentrate solely on the nonresponse error since there is no sampling error. Moreover, the fact that populations of relatively large size (1,000) are generated makes it possible to emphasize the bias of the estimators instead of their variance, since the variance should diminish as the size of the population increases (for a fixed mean response rate).

For each of the 1,000 samples of respondents, we calculated the 7 estimates of the mean described in the two previous sections. We then calculated, for each population, the mean and the variance of these 1,000 estimates, denoted $\bar{\mu}^*$ and $S^2_{\mu^*}$, respectively. Finally, we calculated an estimate of the relative bias (expressed as a percentage), $RB^* = [(\bar{\mu}^* - \mu)/\mu] \times 100\%$, an estimate of the standard error associated with this relative bias, $SE^* = (100/\mu)\sqrt{S^2_{\mu^*}/1\,000}$, and an estimate of the root mean square errors, $RMSE^* = \sqrt{S^2_{\mu^*} + (\bar{\mu}^* - \mu)^2}$.

The results of the simulation study are shown in Table 1. An analysis of this table indicates that, regardless of the error distribution, the relative bias and the mean square error of all the estimators is lower when the correlation between X and Y is greater, which is not surprising.

Table 1
Simulation Results Used to Compare 7 Estimators of the Mean μ

Estimator	$R^2 = 80\%$			$R^2 = 50\%$		
	$RB^*(\%)$	SE^*	$RMSE^*$	$RB^*(\%)$	SE^*	$RMSE^*$
Population with normally distributed errors						
$\hat{\mu}^*_{P,U}$	16.90	0.03	0.84	26.68	0.04	1.33
$\hat{\mu}^*_{P,X}$	5.65	0.02	0.28	18.02	0.03	0.90
$\hat{\mu}^*_{P,Y,ML}$	-0.14	0.03	0.05	1.27	0.10	0.17
$\hat{\mu}^*_{P,Y,ROB}$	1.14	0.03	0.08	10.12	0.06	0.51
$\hat{\mu}^*_{I,X}$	5.50	0.02	0.27	17.74	0.03	0.89
$\hat{\mu}^*_{I,Y,ML}$	0.13	0.03	0.04	1.03	0.07	0.13
$\hat{\mu}^*_{I,Y,ROB}$	0.64	0.03	0.05	7.53	0.06	0.39
Population with exponentially distributed errors						
$\hat{\mu}^*_{P,U}$	17.83	0.04	0.86	26.60	0.05	1.29
$\hat{\mu}^*_{P,X}$	5.44	0.02	0.26	16.06	0.04	0.78
$\hat{\mu}^*_{P,Y,ML}$	-0.54	0.02	0.04	5.18	0.05	0.26
$\hat{\mu}^*_{P,Y,ROB}$	1.31	0.02	0.07	7.43	0.03	0.36
$\hat{\mu}^*_{I,X}$	5.19	0.02	0.25	15.41	0.03	0.75
$\hat{\mu}^*_{I,Y,ML}$	-3.42	0.03	0.17	-25.47	0.05	1.23
$\hat{\mu}^*_{I,Y,ROB}$	0.49	0.02	0.04	4.07	0.03	0.20

An analysis of the relative bias indicates that the method of maximum likelihood provides best results when the errors are normally distributed, followed by the robust estimation method described in section (4.1). Estimators which assume a nonignorable response mechanism have a lower relative bias than those which incorrectly assume an ignorable response mechanism. Among the latter estimators, the most biased is estimator $\hat{\mu}^*_{P,U}$. For a given method, there is generally little difference between the

weighted estimator (2.1) and the estimator that includes imputed values (2.2). However, the latter must be given a slight advantage.

The conclusions in the previous paragraph always apply when errors are exponentially distributed, except that the robust estimation method becomes the best. This observation should not be surprising since the method of maximum likelihood is based on the error normality hypothesis. However, the weighted estimator $\mu^*_{P,Y,ML}$ remains slightly biased, and this is more difficult to explain.

The conclusions drawn from an analysis of the relative bias still apply when analyzing the mean square error. In fact, estimators which are very biased show a strong tendency to having a high mean square error and vice versa.

6. DISCUSSION

When the hypothesis of a nonignorable response mechanism is realistic, and when the hypothesis of error normality for linear regression model (3.2) is justified, using the method of maximum likelihood may be appropriate. However, when the latter hypothesis is not justified, the results of the simulation study described in section 5 show that the robust estimation method presented in this paper is preferable.

Moreover, Beaumont (1999) described the results of another simulation study indicating that the estimation method proposed in this paper is robust with respect to both the error normality hypothesis and model (3.2). As for the method of maximum likelihood, it has been shown to be even more sensitive to the validity of model (3.2) than to the hypothesis of error normality. The latter method should therefore only be used when all the hypotheses associated with models (3.2) and (3.3) are reasonable.

Obviously, all estimators show little bias when nonresponse is very low. Likewise, when the coefficient of correlation between X and Y is very high, all estimators show little bias, except for the estimator which assumes a uniform response mechanism $\mu^*_{P,U}$. In either case, the choice of an estimator should be based on the criterion of simplicity, which favours the estimators in section (3.2), specifically estimator $\mu^*_{I,X}$.

It should be noted that models (3.2) and (3.3) could be complexified according to the nature of the problem. For example, other independent variables could be included in these models. Variable Y could also be categorized using dummy variables, and these dummy variables could be used in model (3.3) instead of variable Y itself.

In this paper, we have dealt only with the problem of the estimation of a mean when the response mechanism is not ignorable. However, the methods described in sections 3 and 4 apply to other types of estimation. For example, weights or imputed values could be used for the estimation of parameters in a given regression.

This paper has attempted to describe a robust estimation method with respect to the hypothesis of error normality for model (3.2), making it possible to reduce the bias due to a nonignorable response mechanism. In some future work, it would be interesting to evaluate simple methods of variance estimation using imputed data and this robust estimation method.

ACKNOWLEDGEMENTS

The author wishes to thank the Small Area and Administrative Data Division of Statistics Canada, which made this work possible. He also wishes to thank Eric Rancourt, the two referees as well as the associate editor for some useful comments which helped improve the quality of this paper.

REFERENCES

BEAUMONT, J.-F., and DEMNATI, A. (1998). Parameter estimation for a finite mixture of distributions for dichotomous longitudinal data: comparing algorithms. *Proceedings, Symposium 98, Longitudinal Analysis for Complex Survey, Statistics Canada*, 191-197.

BEAUMONT, J.-F. (1999). A robust estimation method in the presence of nonignorable nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. (To appear).

D'AGOSTINO, R.B. (1986). Graphical analysis. *Goodness-of-fit Techniques*, (R.B. D'Agostino and M.A. Stephens, Ed.), 7-62. New York: Marcel Dekker.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

GAGNON, F., LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *1996 Proceedings of the Survey Methods Section, Statistical Society of Canada*, 151-156.

GREENLEES, J.S., REECE, W.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.

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

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide*, 2, Version 6, Fourth Edition. Cary, NC: SAS Institute Inc.

ZEGER, S.L., LIANG, K., and ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 137-138
Statistics Canada

An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities

BRUCE D. SPENCER[1]

ABSTRACT

It is common practice to estimate the design effect due to weighting by 1 plus the relative variance of the weights in the sample. This formula has been justified when the selection probabilities are uncorrelated with the variable of interest. An approximation to the design effect is provided to accommodate the situation in which correlation is present.

KEY WORDS: Weighting; Deff; Sampling variance; Complex samples.

1. INTRODUCTION

It is common practice to weight observations in an unequal probability sample by the reciprocals of selection probabilities. The rationale is that failure to use the weights will cause bias if the sampling weights correlate with the variable of interest. A drawback to weighting is an increase in sampling variance when the weights vary excessively in the sample. This increase may be quantified by the design effect. The design effect is the ratio of the variance of the statistic of interest under the design of interest to the variance of the statistic under simple random sampling with the same sample size (Kish 1965). Design effects are important both for approximating standard errors after the sample is in hand and for predicting standard errors ahead of time, which is critical for efficient design of samples.

Kish (1965, 1992) discussed an approximation for the design effect for weighted estimates from unequal probability samples: $1 + rvw$, with rvw defined as the relative variance of the weights in the sample. Thus, if w_i is the weight of unit i in the sample and $\bar{w}$ is the sample mean, $rvw = n^{-1}\sum_{i=1}^{n}(w_i - \bar{w})^2/\bar{w}^2$. Gabler, Haeder, and Lahiri (1999) used a superpopulation model to derive a design effect when clustering is present as well. Their formula, which agrees with design-based results in Kish (1965), reduces to $1 + rvw$ when there is zero intraclass correlation. The $1 + rvw$ approximation for the design effect is based on a model or design in which the weights are uncorrelated with the variable of interest (and hence an unweighted estimate would serve as well or better than the weighted estimate). Here we develop an approximation to the design effect under a model in which correlation may be present. In developing the approximation we do not assume that the population is sampled from a superpopulation. The accuracy of the approximation depends only on the characteristics of the sample design and the population of interest.

For simplicity, we will discuss single-stage unequal probability sampling with replacement. Heuristic extension of the results to sampling without replacement is indicated in section 4.

2. REGRESSION REPRESENTATION OF POPULATION AND SAMPLE DESIGN

Let y_i denote the measurement of interest, P_i the (draw-by-draw) selection probability for a sample of size n, and $w_i = 1/(nP_i)$ the sampling weight for unit i in a population of size N, $1 \le i \le N$. Observe that $\bar{P} = \sum_{i=1}^{N} P_i/N = N^{-1}$. Consider the least-squares population regression line

$$y_i = \alpha + \beta P_i + \varepsilon_i, \tag{1}$$

with $\alpha = \bar{Y} - \beta/N$, $\beta = \sum_{i=1}^{N}(y_i - \bar{Y})(P_i - \bar{P})/\sum_{i=1}^{N}(P_i - \bar{P})^2$, and $\bar{Y} = \sum_{i=1}^{N} y_i/N$. Denote the population variances of the y's, the ε's, the ε^2's, and the w's by $\sigma_y^2, \sigma_\varepsilon^2, \sigma_{\varepsilon^2}^2$, and σ_w^2, with, for example, $\sigma_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/N$. Denote the population correlation between y and P by $\rho_{y,P}$, between ε and w by $\rho_{\varepsilon,w}$, and between ε^2 and w by $\rho_{\varepsilon^2,w}$. It follows from the properties of least-squares, or equivalently from the definitions of α and β, that $\sum_{i=1}^{N}\varepsilon_i P_i = \sum_{i=1}^{N}\varepsilon_i/N = 0$ and $\sigma_\varepsilon^2 = (1 - \rho_{y,P}^2)\sigma_y^2$. If data are available, we can fit the regression representation (1) and estimate α, β, σ, and ρ by, say, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\rho}$.

Let $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ denote the usual weighted estimator of the population total, Y. The variance of $\hat{Y}$ is well-known (Cochran 1977, 253) to be

$$V(\hat{Y}) = n^{-1}\sum_{i=1}^{N} P_i(y_i/P_i - Y)^2. \tag{2}$$

Using the regression formulation (1), we may re-express the variance as

$$V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W} + N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + 2\alpha N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w, \tag{3}$$

where $\bar{W} = \sum_{i=1}^{N} w_i/N$.

This expression does not rest on any assumptions about the fit of the regression model. (See section 5 for derivation).

If the regression model fits well enough so that $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ are zero, then the variance in (3) simplifies to $V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W}$. If simple random sampling

[1] Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, U.S.A.

with replacement had been used, the variance would have been $n^{-1}N^2\sigma_y^2$. Therefore, if $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ are negligible, the design effect is approximately

$$\text{deff} = (1-\rho_{y,P}^2)n\bar{W}/N + (\alpha/\sigma_y)^2(n\bar{W}/N - 1). \quad (4)$$

This approximation does not require that the residuals from the regression are negligible, and it can hold when σ_ε is large. A referee has pointed out that the condition that $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ are negligible may seem unnatural in a model that regresses y on P rather than on $w \propto 1/P$. Note, however, that if we had not only zero correlation between ε and P but also independence, then we would have zero correlation between functions of ε and functions of P, and so $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ would be zero as well.

3. ESTIMATION OF DESIGN EFFECT

To estimate the design effect after the sample is in hand, we may use $1 + rvw$ to estimate $n\bar{W}/N$. To understand the rationale for this, note first that

$$1 + rvw = \frac{n^{-1}\sum_{i=1}^{n} w_i^2}{\bar{w}^2}. \quad (5)$$

The expectation of the numerator is $N\bar{W}/n$. The expectation of $\bar{w}$ is N/n, and so the denominator of (5) may be taken as an estimator of $(N/n)^2$. Dividing the expectation of the numerator by $(N/n)^2$, we obtain $n\bar{W}/N$. Thus the design effect may be estimated from the sample by

$$(1-\hat{\rho}_{y,P}^2)(1 + rvw) + (\hat{\alpha}/\hat{\sigma}_y)^2(rvw). \quad (6)$$

As a special case, note that if we set $\hat{\rho}_{y,P} = 0$, the case of "haphazard weighting" (Kish 1992), then the estimate of the design effect simplifies to

$$1 + rvw + rvw(\hat{\alpha}^2/\hat{\sigma}_y^2). \quad (7)$$

This estimate is close to Kish's approximation when $\hat{\alpha}/\hat{\sigma}_y$ is near zero.

4. SAMPLING WITHOUT REPLACEMENT

To derive the exact design effect for sampling without replacement would be more complex, as it would require consideration of joint selection probabilities for pairs of units. A heuristic extension of the results is easy, however. Recall that the ratio of the variance of a sample mean under simple random sampling without replacement to the variance under with-replacement sampling is approximately $(1 - n/N)$.

The results we have derived for the design effect will apply to single-stage unequal probability samples of n units without replacement if the variance of the Horvitz-Thompson estimator of the total is approximately $(1-n/N)$ times the variance in (2), with P_i taken as n^{-1} times the overall selection probability for unit i (Särndal, Swensson, and Wretman 1992, 154).

5. DERIVATION OF VARIANCE FORMULA (3)

From (2) we have $V(\hat{Y}) = n^{-1}(\sum_{i=1}^{N} y_i^2/P_i - Y^2)$. Next, note that (1) implies that

$$Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2 \quad (8)$$

and

$$\begin{aligned}\sum_{i=1}^{N} y_i^2/P_i &= \sum_{i=1}^{N}[\alpha^2/P_i + \beta^2 P_i + \varepsilon_i^2/P_i + 2\alpha\beta + 2\alpha\varepsilon_i/P_i + 2\beta\varepsilon_i] \\ &= \alpha^2\sum_{i=1}^{N} P_i^{-1} + \beta^2 + \sum_{i=1}^{N}\varepsilon_i^2/P_i + 2N\alpha\beta + 2\alpha\sum_{i=1}^{N}\varepsilon_i/P_i \\ &= \alpha^2 n\sum_{i=1}^{N} w_i + \beta^2 + n\sum_{i=1}^{N}\varepsilon_i^2 w_i + 2N\alpha\beta + 2\alpha n\sum_{i=1}^{N}\varepsilon_i w_i. \quad (9)\end{aligned}$$

Subtracting (8) from (9) and dividing by n yields

$$V(\hat{Y}) = \alpha^2\left(\sum_{i=1}^{N} w_i - N^2/n\right) + \sum_{i=1}^{N}\varepsilon_i^2 w_i + 2\alpha\sum_{i=1}^{N}\varepsilon_i w_i.$$

To obtain (3), note that

$$\begin{aligned}\sum_{i=1}^{N}\varepsilon_i^2 w_i &= N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + N\bar{W}\sigma_\varepsilon^2 \\ &= N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + (1-\rho_{y,P}^2)\sigma_y^2 N\bar{W}\end{aligned}$$

and

$$\sum_{i=1}^{N}\varepsilon_i w_i = N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w.$$

REFERENCES

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

GABLER, S., HAEDER, S., and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 1, 105-106.

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

KISH, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 2, 183-200.

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

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 139-152
Statistics Canada

On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data

PAUL P. BIEMER and JOHN M. BUSHERY[1]

ABSTRACT

The primary goal of this research is to investigate the validity of Markov latent class analysis (MLCA) estimates of labor force classification error and to evaluate the efficacy of MLC analysis as an alternative to traditional methods for evaluating data quality. We analyze interview data from the Current Population Survey (CPS) for the first three months of each of three years – 1993, 1995, and 1996 – and conduct an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provides another approach for estimating CPS classification error that, when compared with the MLC estimates, helps to address the validity of the MLCA approach. Five dimensions of MLCA validity are addressed: (a) model diagnostics, (b) model goodness of fit across three years of CPS, (c) agreement between the model and test-retest reinterview estimates of response probabilities, (d) agreement between the model and test-retest reinterview estimates of inconsistency, and (e) the plausibility of patterns of classification error. In addition, we consider the robustness of the MLCA estimates to violations in the Markov assumption. Our analyses provides no evidence to question the validity of the MLC approach. The method performed well in all five validity tests.

KEY WORDS: Panel surveys; Nonsampling error; Unemployment; Data quality.

1. INTRODUCTION

The Current Population Survey (CPS) is a household sample survey conducted monthly by the U.S. Bureau of the Census to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. National estimates from the CPS of the size, composition, and changes in the composition of the labor force are published each month by the U.S. Bureau of Labor Statistics in Employment and Earnings. The CPS labor force estimates comprise one of the Nation's key economic indicators; since 1942, the Federal government has used the CPS data series to monitor month-to-month and year-to-year changes in labor force participation.

Given the importance of the CPS data series to public policy, there have been numerous evaluations of the accuracy of the data. For example, since the early 1950s, the Census Bureau has conducted the CPS Reinterview Program to evaluate the quality of the labor force data. The program involves drawing a small subsample (less than 5 percent) of the CPS respondents and re-asking some of the questions asked in the original interview – particularly the labor force questions. Until 1994, about one fourth of the sample received an unreconciled reinterview and three fourths received a reconciled reinterview. The reconciled reinterview component, which was used primarily for interview quality control purposes, was discontinued in 1994 due to concerns about the quality of the data. However, the unreconciled reinterview continues today and is used to estimate the test-retest reliability (or response consistency). Forsman and Schreiner (1991) provide a detailed description of the CPS Reinterview Program.

Several papers prepared by researchers outside the Census Bureau analyze the CPS Reinterview Program data to estimate the classification error in the CPS (*cf.* Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Chua and Fuller 1987; Poterba and Summers 1986; Abowd and Zellner 1985). Recently, Poterba and Summers (1995) used data from the CPS Reinterview Program to estimate the CPS classification error rates and to evaluate the impact of classification error on labor market transition rates. As in the 1986 paper, their more recent analysis is based on the assumption that the CPS reinterview reconciliation process yields data which may be considered as the truth. Abowd and Zellner (1985) took similar approach.

Several authors (*viz.*, Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Forsman and Schreiner 1991; Schreiner 1980) question the assumption that reconciled reinterview yields true values. They provide considerable evidence that the reinterview data are subject to substantial classification errors. In fact, this realization was responsible for the Census Bureau's decision to eliminate the reconciled reinterview portion of the CPS Reinterview Program in 1994.

As an alternative to the infallibility assumption, Chua and Fuller (1987) and Fuller and Chua (1985) apply a type of latent structure model to the CPS reconciled reinterview data to estimate the CPS response probabilities. For model identifiability, they impose tight restrictions on the response probabilities, forcing the bias due to classification error to be zero for both interview and reinterview. In addition, they assume independent classification errors for the interview and reinterview (referred to as the ICE assumption in the literature) and across the months in sample. The ICE

[1] Paul P. Biemer, Research Triangle Institute, Research Triangle Park, NC 27709; John M. Bushery, Bureau of Transportation Statistics, Washington, DC 20590, U.S.A.

assumption is a limitation of their analysis because evidence in the literature suggests that the assumption may not hold for the CPS (see, for example, O'Muircheartaigh 1991, and Singh and Rao 1995). Consequently, response probabilities estimated using the Chua and Fuller approach may be biased.

Sinclair and Gastwirth (1996) and Sinclair and Gastwirth (1998) apply a latent class modeling approach to the CPS interview-reinterview data using model restrictions originally proposed by Hui and Walter (1980) for medical diagnostic testing. Using the interview-reinterview data cross-classified by sex, Sinclair and Gastwirth assume that classification error probabilities are equal for males and females while labor force participation rates differ for these groups. Since the model parameters consume all the available degrees of freedom for parameter estimation, no residual degrees of freedom are available to test model lack-of-fit. Consequently, their analysis does not directly address whether these model assumptions hold for the CPS data.

In an examination of the determinants of rotation group bias, Shockey (1988) also applies latent class analysis to the CPS. His analysis suggests that the rotation group bias problem first reported by Bailar (1975) may be caused by response error arising from the interview administration. Shockey did not use reinterview data but rather relied on confirmatory factor analytic methods to support his claims. The sizes of his error rates were much larger than those reported by other authors which may be an indication of model bias. Unfortunately, like Sinclair and Gastwirth, Shockey's data set is not adequate to test fully the assumptions of the model he used.

The method of Markov latent class analysis, a promising approach for estimating the classification error in panel survey data, previously has not been applied to the CPS. This method takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data. The MLCA model is really a combination of two models: a latent Markov chain model representing the month to month transitions among the true labor force classifications and a classification error model representing the deviations from the true and observed labor classifications.

Because MLCA takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data, it does not require external, infallible measurements or remeasurements obtained by reinterview methods. In that regard, the method offers some advantages over both the Census Bureau's traditional methods and the methods of Chua and Fuller, Abowd and Zellner, Porterba and Summers, and Sinclair and Gastwirth for evaluating survey data quality in surveys. In many panel surveys, reinterviews are not feasible due to budget constraints, field work complexity, and respondent burden. MLCA may be the only way to assess the measurement error in these surveys. For panel surveys, such as the CPS, where reinterview data are available, the reinterview and MLCA methods offer alternative analytical approaches for evaluating classification error. For example, as in the present analysis, MLCA can be used to model and test the traditional reinterview analysis assumptions. Further, MLCA analysis provides a statistical framework for combining the panel data and reinterview data to obtain even more information about classification error (van de Pol and Langeheine 1997).

Another advantage of MLCA is the potential for incorporating the entire panel data set into the estimates of classification error rather than only the relatively small sample selected for reinterview. As a result, a number of data quality issues for panel surveys that previously could not be explored for lack of data may now be tractable.

This paper reports our findings regarding the utility of the MLCA modeling approach for evaluating labor force classification error in the CPS. Software for fitting a wide variety of MLCA and other latent class models is available from several sources. The software employed in our analysis is ℓEM (Vermunt 1997), which can fit a large class of log-linear models with or without latent variables. The flexibility and generality of this software allow the measurement error analyst to test a considerable range of classification error models and to explore hypotheses regarding the causes and correlates of classification error.

In the next section, we describe the MLCA model and estimation methodology and its theoretical underpinnings. In section 3, we develop the MLCA methodology for the CPS application, fit a series of models to the CPS, and examine the fit of these models. In this section, we also produce estimates of classification error based upon the best MLCA model. In section 4, we conduct a number of tests of the validity of the MLCA estimates including a comparison of the MLCA estimates with those from new interview-reinterview analysis. Finally, in section 5, we summarize our findings and make recommendations regarding the utility of the MLCA method for future evaluations of labor force classification error.

2. MARKOV LATENT CLASS ANALYSIS FOR THREE TIME PERIODS

Markov latent class models were first proposed by Wiggins (1973) and refined by Poulsen (1982). Van de Pol and de Leeuw (1986) established conditions under which the model is identifiable and gave other conditions of estimability of the model parameters. In this section, we develop the MLCA model in the context of the CPS and suggest other applications and its generalizations.

Let the CPS target population be divided into L groups (such as age, race, or sex groups) and let the variable G be the label for group membership. For example, $G_i = 1$ if the i-th population member is in group 1, $G_i = 2$ for group 2 and so on. Let X_{gi}, Y_{gi}, and Z_{gk} denote the true labor force classifications for the i-th person in group

$G = g$(for $g = 1, ..., L$ and $i = 1, ..., n_g$) where X_{gi} is defined as

$$X_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is employed in time period 1} \\ 2 & \text{if person } (g, i) \text{ is unemployed in time period 1} \\ 3 & \text{if person } (g, i) \text{ is not in the labor force in time period 1} \end{cases}$$

with analogous definitions for Y_{gi} and Z_{gi} for periods 2 and 3 respectively. Let $\pi_{x,y,z|g}$ denote $\Pr(X = x, Y = y, Z = z|G = g)$, let $\pi_{y|g,x}$ denote $\Pr(Y = y|X = x, G = g)$ and let $\pi_{z|g,y,x}$ denote $\Pr(Z = z|Y = y, X = x, G = g)$. Then, the probability that an individual in group g has labor status x in period 1, y in period 2, and z in period 3 is

$$\pi_{x,y,z|g} = \pi_{x|g}\,\pi_{y|g,x}\,\pi_{z|g,x,y}. \qquad (1)$$

Finally, under the first order Markov assumption, a necessary condition for model identifiability (see Van de Pol and de Leeuw 1986), we assume

$$\pi_{z|g,x,y} = \pi_{z|g,y} \qquad (2)$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status, once the period 2 status is known. An alternate interpretation is that the current status, given the prior period's status, does not depend upon the prior period's transition.

One can conceive of a number of scenarios where the Markov assumption may not hold for monthly labor force status. The assumption would be violated, for example, if individuals who are unemployed in period 2 are more likely to be unemployed in period 3, given they were also unemployed in period 1. The group of people unemployed in period 2 and period 1 probably includes a higher proportion of chronically unemployed people than the group unemployed in period 2 but not in period 1. That group (unemployed period 2, not period 1) likely contains a higher proportion of people temporarily out of work while changing jobs.

However, the validity of this assumption cannot be adequately explored using the observed data because the data are distorted to some unknown extent by the presence of classification errors. At least two methods for assessing the validity of the Markov assumption for panel data are available. Van de Pol and de Leeuw (1986) suggest a method based upon four waves of panel data that substitutes a second order Markov restriction for the first order restriction in (2). Another method, suggested by van de Pol and Langeheine (1997), uses a combination of labor force panel data and the reinterview data at each time period. Neither of these methods was employed in this paper to test the MLCA assumption directly. Instead, we assessed the overall validity of the MLCA estimates using the methods discussed in section 3.2 below. In section 3.6 we provide some results from a simulation study to illustrate the robustness of the MLCA estimates of classification error to violations of the Markov assumption.

Now, consider the observed labor force classifications from the CPS denoted by A_{gi}, B_{gi}, and C_{gi} for periods 1, 2, and 3, respectively, where

$$A_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is classified as employed in time period 1} \\ 2 & \text{if person } (g, i) \text{ is classified as unemployed in time period 1} \\ 3 & \text{if person } (g, i) \text{ is classified as NLF in time period 1} \end{cases}$$

with analogous definitions for the response indicators, B_{gi}, and C_{gi} for periods 2 and 3, respectively. Using an extension of the notation established above, we denote the response probabilities in each of these classifications as $\pi_{a|g,x} = \Pr(A = a|G = g, X = x)$, with analogous definitions for $\pi_{b|g,y}$ and $\pi_{c|g,z}$. Thus, $\pi_{a=1|g,x=2}$ is the probability that the CPS classifies a person in group g as employed ($A = 1$) when the true status is unemployed ($X = 2$). Likewise, $\pi_{a=2|g,x=2}$ is the probability that the CPS correctly classifies a person in group g as unemployed.

Finally, we assume

$$\pi_{a,b,c|g,x,y,z} = \pi_{a|gx}\,\pi_{b|gy}\,\pi_{c|gz} \qquad (3)$$

or that classification error in the observed labor forces status is independent across the three months. This assumption, referred to as the local independence assumption, has been investigated for the CPS by Meyers (1988) in his review of the Abowd and Zellner (1985) estimation approach. Meyers concluded that the assumption "seems a reasonable approximation." Singh and Rao (1995), who studied the robustness of the assumption under a number of labor force population scenarios, reached a similar conclusion. Van de Pol and Langeheine (1997) modeled the joint distribution of panel data and reinterview data using latent class models to test for local independence for various types of labor force transitions. They found some evidence that people who change labor force status have lower reliability than those who do not, however the effect was quite small. Therefore, we shall also assume (3) without attempting any further investigation of its validity in this paper.

The CPS labor force classifications for each month of the first quarter of the year are the outcome variables in our analysis. Let A, B, and C denote the observed classifications and let X, Y, and Z denote the (unobserved) true classifications for January, February, and March, respectively. Let G denote some grouping (or stratification) variable to be defined later in the analysis. Under these assumptions, we can write the probability for classifying a CPS sample member in cell (g, a, b, c) of the $GABC$ table as follows:

$$\pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{a|g,x} \pi_{y|x,g} \pi_{b|y,g} \pi_{z|g,y} \pi_{c|g,z} \quad (4)$$

Extensions to more than one grouping variable are straightforward.

Under multinomial sampling, the likelihood function for the $GABC$ table is

$$\Pr(GABC) = k \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{gabc}} \quad (5)$$

where k is the multinomial constant and Π denotes the product of the terms over the subscripts g, a, b, and c. Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provides the formula for applying the E-M algorithm to estimate the parameters of this model and describes the conditions for their estimability. The ℓEM software, applied to the CPS data sets in the next section, implements these methods.

3. APPLICATION TO THE CPS

3.1 Notation

Part of our evaluation of the MLCA approach will compare the MLCA estimates of classification error with estimates derived from the analysis of interview-reinterview data. Using the notation in the previous section, let A and A' denote the labor force classification for the original and reinterview, respectively, and define $\pi_a = \Pr(A = a)$ and $\pi_{a'} = \Pr(A' = a')$. Let AA' denote the observed interview-reinterview $K \times K$ cross-classification table and let $\boldsymbol{\pi}^{AA'|X}$ denote the $K \times K$ matrix of cell probabilities, $\Pr(A = a, A' = a' | X = x)$. If we assume that $\pi_{aa'} = \Pr(A = a, A' = a' | X = x) = \pi^2_{a|x}$, referred to in the literature as the assumption of parallel measures (Bohrnstedt 1983), then

$$\boldsymbol{\pi}^{AA'|X} = \boldsymbol{\pi}^{A|X} (\boldsymbol{\pi}^{A|X})^T \quad (6)$$

where $(\boldsymbol{\pi}^{A|X})^T$ denotes the transpose of vector of conditional probabilities, $\boldsymbol{\pi}^{A|X}$.

Let $\boldsymbol{\pi}^X$ denote the K-vector of true classification probabilities. Then

$$\boldsymbol{\pi}^{AA'} = \boldsymbol{\pi}^{AA'|X} \boldsymbol{\pi}^X \quad (7)$$

i.e., the probability of the observed interview-reinterview classification table, $\boldsymbol{\pi}^{AA'}$, is equal to the product of the matrix of conditional response probabilities, $\boldsymbol{\pi}^{AA'|X}$, and the vector of true classification probabilities, $\boldsymbol{\pi}^X$.

As described in the previous section, the MLCA of the CPS longitudinal data will provide maximum likelihood estimates of $\boldsymbol{\pi}^{A|X}$ and $\boldsymbol{\pi}^X$, allowing the estimation of $\boldsymbol{\pi}^{AA'}$ via (6) and (7). We can estimate the test-retest reliability, R, for any labor force category by applying the usual estimation methods (see, for example, Bohrnstedt 1983) to this estimate of $\boldsymbol{\pi}^{AA'}$. For our analysis, we compute the index of inconsistency, $I = 1 - R$, which is the traditional reliability measure for CPS labor force data (see U.S. Bureau of the Census 1985). Let I_a, denote the index of inconsistency for category $A = a$. Then an estimator of I_a is

$$\frac{gdr}{2\hat{\pi}_a(1 - \hat{\pi}_a)} \quad (8)$$

where gdr is the gross difference rate defined by

$$gdr_a = 2 \sum_{a \neq a'} \hat{\pi}_{a,a'} \quad (9)$$

and where $\hat{\pi}_a$ and $\hat{\pi}_{a,a'}$ denote latent class estimates of π_a and $\pi_{a,a'}$, respectively.

U.S. Bureau of the Census (1985, 88-91) provides the formulas for standard errors as well as an aggregate measure of inconsistency for all K categories combined, referred to as the aggregate index of inconsistency, I_{AG}. The aggregate index is a question-level measure of unreliability equal to $1 - \kappa$ (Hess, Singer and Bushery 2000) where κ is Cohen's kappa reliability measure (Cohen 1960) and is a weighted average of the category-level indexes.

Finally, given an estimate of $\boldsymbol{\pi}^X$ we can estimate the K-vector of measurement biases, denoted by $\boldsymbol{\beta}_A$, associated with the K categories of A using the identity

$$\boldsymbol{\beta}_A = \boldsymbol{\pi}^A - \boldsymbol{\pi}^X. \quad (10)$$

3.2 Assessing the Validity of the MLCA Methodology

The primary objective of this paper is to assess the validity of the MLCA approach. Previous research in the measurement of CPS classification error has not fully addressed the validity of the estimation approaches used (Meyers 1988). We hope to determine whether the MLCA approach is informative and useful for studying classification error in the CPS. In particular, we aim to determine whether the model estimates of error probabilities, $\boldsymbol{\pi}^{A|X}$, reflect the actual levels of error in the CPS labor force classifications. Unfortunately, for the reasons mentioned previously, no generally accepted gold standard exists for assessing the accuracy of the CPS (see, for example, Sinclair and Gastwirth 1996, 1998, Biemer and Forsman 1992, and Schreiner 1980). Consequently, estimating the bias of MLCA estimates is not possible.

In what follows, we will investigate the validity of the MLCA estimates of CPS classification error using five criteria:

1. **Model diagnostics.** A necessary condition for model validity is that the model is plausible (*i.e.*, the assumptions are reasonable and are consistent with reality) and fits the data adequately. We use the traditional chi-square goodness of fit criteria and other diagnostic measures of model fit to assess the adequacy of the model specification and the degree to which the data are consistent with the model.

2. **Model Goodness of Fit Across Years of CPS.** An often-used technique for model validation is to assess the fit of the model for data that are independent of the data used for model building (see, for example, Kleinbaum, Kupper and Muller 1988, 330). This method is useful for avoiding model over-parameterization and data-driven (rather than theory-driven) model selection. In the present study, fitting the same model to data for each year separately is a form of this independent model verification technique. Model agreement across the years would tend to support the validity of the model structure. This method has a difficulty in the present application. After 1993, the CPS paper and pencil questionnaire was redesigned for Computer Assisted Personal Interview (CAPI) administration, so the magnitudes of the response errors may have changed after 1993. However, if the primary sources of response error in the CPS have not changed with the redesign, a model structure that adequately describes the error for 1993 should also describe the error for 1995 and 1996.

3. **Agreement of the MLCA Estimates and the Hui-Water Test-Retest Estimates of Response Probabilities.** The Hui-Walter (H-W) method (Hui and Walter 1980) for estimating CPS response probabilities uses unreconciled reinterview data (Sinclair and Gastwirth 1996; 1998). Although the MLCA and H-W methods both use latent class models, the model assumptions are very different. For example, the H-W method does not require the Markov assumption for model identifiability. Further, in this research, the data inputs to the H-W method are independent of those used for the MLCA method. Close agreement between the two sets of estimates supports the validity of both methods, while poor agreement suggests that at least one of the approaches is not valid. Strong agreement between the MLCA and H-W estimates also lend some assurance that the MLCA estimates of response probabilities are relatively robust to possible violations of the Markov assumption.

4. **Agreement of Model and Test-Retest Estimates of the Index of Inconsistency.** This criterion is similar to Criterion 3 because it compares estimates derived from MLCA with estimates based upon unreconciled reinterview data. However, this analysis does not rely on the validity of the Hui-Walter estimation methodology to assess MLCA estimation validity. Instead we use the MLCA estimates of classification error to compute estimates of the index of inconsistency using (7) to (9). We compare these estimates of reliability directly to the estimates of reliability from the CPS Reinterview Program, obtained from unreconciled reinterview data. Good agreement between the Reinterview and MLCA estimates supports the validity of both methods, while poor agreement suggests that at least one of the approaches is not valid.

5. **Plausibility of Patterns of Classification Error.** Finally, the plausibility (or face validity) of the response probability estimates can also provide a test of validity. For example, it seems implausible that proxy responses to labor force questions should be more accurate than self-responses. Other patterns of classification error can also be reviewed and evaluated for plausibility. To the extent that the model estimates seem plausible, the face validity of the estimates is supported.

In the next section, we discuss our MLCA modeling results in the context of these criteria for validity. We begin with a description of the CPS data sets and the results of the model selection process.

3.3 The CPS Data Sets

In 1994, in conjunction with the implementation of computer assisted personal interviewing (CAPI), the CPS underwent a major redesign and a restructuring of the questions used to determine labor force status. Rothgeb (1994) provides a description of the CPS redesign. As a result of these improvements, we expect to see a difference (specifically a reduction) in classification error for the post-1994 CPS relative to 1993. Although not a primary objective of this research, we compared the error in the CPS before and after the redesign. We tested the MLCA approach for three years of the CPS – 1993, 1995, and 1996 – because the CPS unreconciled reinterview data were readily available for these time periods.

The CPS households are interviewed for four consecutive months, drop out of the survey for eight months, and then re-enter to be interviewed for a second series of four consecutive months. MLCA requires at least three consecutive interviews for identifiability of the model parameters. We had a choice of data sets which included all persons interviewed in three or four consecutive months of the CPS. Since using four months of data would reduce the sample size for the analysis by half, we chose to focus the analysis on three consecutive months – January, February, and March – for all three years of data. Nonresponse cases and cases where the whole household changed in one or more of the three months were excluded from the analysis.

The simplest MLCA model specifies that the response probabilities, $\pi_{a|x}$, $\pi_{b|y}$, and $\pi_{c|z}$, and the transition probabilities, $\pi_{y|x}$, $\pi_{z|y}$ are the same for all persons in the target population (referred to as homogeneity). However, our preliminary analysis (Biemer, Bushery and Flanagan 1997) indicated that response and transition probabilities where not homogeneous. To account for this heterogeneity, we explored a number of covariates and stratification variables for inclusion in the models, including: gender, education, mode of interview, proxy/self-response, and race. Of the

those considered, a variable derived from the CPS proxy/self response indicator best accounted for population heterogeneity. This variable, denoted by P, is defined as follows:

$$P = \begin{cases} 1 & \text{if all three interviews are conducted by self-response (SELF)} \\ 2 & \text{if two of the three interviews are conducted by self-response (MOSTLY SELF)} \\ 3 & \text{if two of the three interviews are conducted by proxy response (MOSTLY PROXY)} \\ 4 & \text{if all three interviews are conducted by proxy response (PROXY)} \end{cases}$$

Note, we now use P to represent the grouping variable, in place of G, which we used in section 2. Based upon previous research (for example, O'Muircheartaigh 1991), we expect that the Self group (P=1) to have less classification error than the Proxy group (P=4). We test this hypothesis as part of the estimate plausibility criterion (criterion 4 above).

The sample sizes for the three data sets used in our analysis are

1993: 45,291 persons
1995: 49,347 persons
1996: 41,751 persons

For 1993, approximately one-third of the sample is in the Self group, approximately one- fourth in the Proxy group, and the remaining sample members are distributed approximately equally between the Mostly Self and Mostly Proxy groups. For 1995 and 1996, slightly more sample members (one-third rather than one-fourth) are in the Proxy group.

3.4 Fitting the MLCA Models

To fit an MLCA model with a single grouping variable, P, the input data set was a 4×3×3×3 table of cell counts defined by the cross-classification of P×A×B×C, where A, B, and C are the labor force classifications for January, February, and March, respectively.

The ℓEM software and other software packages for fitting MLCA models assume simple random sampling, so the complex survey design of the CPS cannot be modeled exactly. It is possible to account for the unequal probability sampling structure of the CPS through the use of weighted and rescaled cell counts rather than the raw cell totals (Clogg and Eliason 1985). However, using unweighted data for the MLCA analysis affords two important advantages. First, we can compare the MLCA estimates with estimates from the previously cited studies on CPS classification error, all of which used unweighted data. Second, the CPS reinterview data used to assess Criteria 3 and 4 are unweighted and weights are not available. Consequently, at least part of the analysis requires unweighted data; using weighted data for the other criteria could produce spurious inconsistencies in the results.

To investigate the validity of inferences to the total population using unweighted analysis, we estimated classification errors from both weighted and unweighted data and observed that the classification error estimates expressed as proportions were virtually identical, differing only at the third decimal place. Thus, the results we report below using unweighted cell counts are appropriate for inference beyond the CPS sample to the total population.

Another consideration in using unweighted analysis is the estimation of standard errors. Since they are computed using simple random sampling assumptions, the ℓEM standard error estimates may be understated as a result of ignoring the clustering effects in the CPS sample. To approximately account for this, we can multiply the ℓEM variances by a design effect computed from the CPS labor force estimates. U.S. Bureau of the Census (2000, 14-9) indicates that the design effects for the CPS labor force estimates do not exceed 1.3 and thus multiplying the ℓEM standard errors by (1.3)½ should inflate the standard errors sufficiently to account for clustering. An equivalent approach is to use a 3 percent rather than a 5 percent level of significance in declaring the difference between two estimates to be statistically significant. This latter strategy will be employed in the forthcoming analysis as appropriate. We believe this produces a conservative test since the CPS design effect reflects the increase in variance due to both sample clustering and unequal weighting, while only clustering effects are present our unweighted estimates.

Table 1 shows the results of fitting a sequence of increasingly complex MLCA models for each of the three data sets. The Base Model is the simplest MLCA model and specifies that transition probabilities and response probabilities are homogeneous (*i.e.*, do not differ by group, P) and stationary (*i.e.*, are the same for all three months). This model may be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|g} \pi^3_{a|x} \pi^2_{y|x}, \qquad (11)$$

which is obtained from (4) by imposing the constraints

$$\pi_{z|yp} = \pi_{y|xp} = \pi_{y|x} \qquad (12)$$

and

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} = \pi_{a|x} \qquad (13)$$

for all p.

For Model 1 we relax constraint (12) to

$$\pi_{z|yp} = \pi_{y|xp} \text{ for } p = 1,...,4 \qquad (14)$$

and thus allow transitions from January to February and February to March to vary by Self/Proxy Group, P. For Model 2, we further relax constraint (12) to

$$\pi_{y|xp} = \pi_{y|x} \text{ and } \pi_{z|yp} = \pi_{z|y} \qquad (15)$$

Table 1
Model Diagnostics for Alternative MLCA Models by Year

1993 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	645	0	-320	0.048
Model 1: Nonhomogeneous transitions	84	23	632	0	-269	0.047
Model 2: Non-stationary transitions	66	41	99	0.006	-609	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	64	0.016	-386	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	23	0.501	-234	0
1995 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	697	0	-275	0.044
Model 1: Nonhomogeneous transitions	84	23	668	0	-240	0.043
Model 2: Non-stationary transitions	66	41	146	0	-567	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82	0	-372	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	25	0.41	-234	0
1996 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	632	0	-325	0.045
Model 1: Nonhomogeneous transitions	84	23	585	0	-308	0.044
Model 2: Non-stationary transitions	66	41	159	0	-543	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82.6	0	-364	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	39.3	0.026	-216	0

[1] Note that "npar" refers to the number of parameters in the model

for all p. Model 3 relaxes both the homogeneity and stationarity constraints for transition probabilities so that $\pi_{y|xp} \neq \pi_{z|yp}$. Thus this model allows transition probabilities to vary by group and by month. However, response probabilities are still constrained to be equal across groups and months.

Model 4 is the most general, identifiable model we considered. Model 4 allows the January-February and February-March transition probabilities to vary independently across the four proxy/self groups. This model further species that the response probabilities are the same for January, February, and March, but may vary across the four proxy/self groups. We obtained this model from Model 3 by relaxing the constraints specifying homogeneous response probabilities; *i.e.*, by relaxing constraint (13) to

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} \qquad (16)$$

for all p. Under these constraints, (4) can be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|p} \pi_{y|xp} \pi_{z|py} (\pi_{a|p,x})^3.$$

In Table 1, we show the basic fit statistics for all five models for all three years. Column 4 of the table provides L^2, the usual likelihood ratio chi-square statistic (see Agresti 1990, 48), and column 5 the corresponding p-value. A p-value of 0.05 or greater is the usual criterion for adequate model fit. However, due to the large sample sizes in our analysis, requiring a p-value this large could result in model over fitting. We consider a p-value as small as 0.01 to be acceptable. The BIC measure in the table is defined as

$$\text{BIC} = L^2 - (\log N)df$$

where N is the total sample size and df is the degrees of freedom for the model. The BIC essentially summarizes the tradeoff between model fit (L2) and model parsimony (df). Since small values of the BIC are favorable, we will regard the model with the smallest BIC as best with respect to goodness of fit and parsimony. Liu and Dayton (1997) discuss this approach for latent class models.

Finally, the dissimilarity index (d) is the proportion of observations that would have to change cells for the model to fit perfectly. As rule of thumb, models having $d \leq 0.05$ (*i.e.*, 5 percent model error) are considered to fit the data well (Vermunt 1997).

For each year of data, Model 4 is the only model to provide an acceptable fit when the p-value criterion is

considered. Model 4 is also plausible from a response theory perspective since it postulates that classification error varies by self/proxy group. This, as we have said, is consistent with the survey methods literature (see for example, O'Muircheartaigh 1991 and Moore 1988). The dissimilarity index, d, for the model is 0.3 percent, which indicates a very good model fit. Thus, we use Model 4 to generate the estimates of labor force classification error.

3.5 Estimation of Classification Error

Table 2 shows estimates of the response probabilities from Model 4 for Employed, Unemployed, and Not in the Labor Force. For the true Employed and those truly Not in the Labor Force, the probability of a correct response is quite high: at least 98 percent for Employed and 97 percent for Not in the Labor Force. However, for the true Unemployed, the probability of a correct response varies across years and groups from approximately 68 percent to approximately 86 percent. As expected, the Self group has the highest probability of a correct response (statistically significant for 1993 and 1996). The Mostly Self group also displays a tendency for a higher probability of a correct response than the Mostly Proxy group, but the difference is not statistically significant.

A surprising result from Table 2 is the direction of the difference in reporting accuracy for the true unemployed across the three years. Recall that the CPS interview questionnaire was redesigned in 1994 to increase reporting accuracy. However, these results suggest that reporting accuracy is higher for 1993 (the year prior to the major redesign) than for 1995 and 1996 (the years following the redesign). In 1993, the probability of a correct response was 81.8 percent, compared with 76.1 percent and 74.4 percent for 1995 and 1996, respectively (p-value <0.001)). This effect may be a consequence the redesign or may reflect actual changes in the population, or both. Investigations are currently underway at the Census Bureau to understand the causes underlying these results.

Table 2
Estimated Labor Force Classification Probabilities by Group and Year
(Standard Errors are in parentheses)

Observed		True Employed			True Unemployed			True NLF		
	Group	1993	1995	1996	1993	1995	1996	1993	1995	1996
Emp	Self	98.90 (0.12)	98.98 (0.12)	99.03 (0.13)	4.24 (0.96)	5.66 (1.28)	5.78 (1.52)	1.05 (0.13)	0.67 (0.10)	0.92 (0.13)
	Mostly Self	98.91 (0.18)	98.61 (0.21)	98.86 (0.22)	8.11 (1.89)	7.66 (2.50)	12.20 (2.96)	1.73 (0.27)	1.59 (0.30)	0.93 (0.24)
	Mostly Proxy	98.62 (0.18)	98.56 (0.22)	98.76 (0.20)	11.19 (1.96)	12.86 (2.97)	10.18 (2.73)	1.68 (0.27)	0.95 (0.25)	1.50 (0.30)
	Proxy	98.66 (0.16)	98.63 (0.13)	98.66 (0.15)	6.60 (1.43)	7.77 (1.44)	8.30 (1.62)	1.56 (0.20)	1.50 (0.17)	1.30 (0.18)
	Total	98.77 (0.11)	98.73 (0.11)	98.83 (0.11)	7.06 (0.70)	7.86 (0.90)	8.57 (1.00)	1.41 (0.11)	1.11 (0.11)	1.13 (0.11)
Unemp	Self	0.33 (0.08)	0.39 (0.08)	0.41 (0.09)	85.92 (1.62)	80.52 (2.04)	79.66 (2.44)	0.48 (0.11)	0.35 (0.10)	0.40 (0.12)
	Mostly Self	0.28 (0.12)	0.50 (0.13)	0.28 (0.15)	82.68 (2.48)	71.71 (3.40)	73.60 (3.71)	1.03 (0.23)	0.91 (0.21)	0.79 (0.28)
	Mostly Proxy	0.37 (0.11)	0.54 (0.15)	0.37 (0.13)	76.54 (2.55)	68.09 (3.37)	72.74 (3.70)	0.81 (0.22)	0.90 (0.28)	1.53 (0.26)
	Proxy	0.34 (0.10)	0.55 (0.09)	0.39 (0.09)	80.09 (2.19)	77.12 (2.18)	71.63 (2.49)	0.95 (0.19)	0.94 (0.17)	1.18 (0.20)
	Total	0.34 (0.11)	0.49 (0.11)	0.37 (0.11)	81.81 (0.90)	76.09 (1.21)	74.42 (1.21)	0.75 (0.11)	0.69 (0.11)	0.87 (0.11)
NLF	Self	0.77 (0.10)	0.63 (0.08)	0.55 (0.09)	9.84 (1.39)	13.82 (1.76)	14.56 (2.11)	98.47 (0.17)	98.98 (0.14)	98.68 (0.18)
	Mostly Self	0.81 (0.13)	0.89 (0.15)	0.86 (0.16)	9.21 (1.83)	20.63 (2.94)	14.20 (2.94)	97.24 (0.35)	97.50 (0.36)	98.28 (0.37)
	Mostly Proxy	1.01 (0.14)	0.90 (0.15)	0.87 (0.15)	12.27 (1.97)	19.05 (2.58)	17.08 (3.16)	97.52 (0.35)	98.15 (0.38)	96.96 (0.40)
	Proxy	0.10 (0.13)	0.82 (0.10)	0.95 (0.12)	13.31 (1.83)	15.11 (1.86)	20.07 (2.22)	97.49 (0.27)	97.56 (0.24)	97.52 (0.27)
	Total	0.89 (0.11)	0.78 (0.11)	0.79 (0.11)	11.13 (0.90)	16.04 (1.21)	17.00 (1.21)	97.84 (0.11)	98.20 (0.11)	98.00 (0.11)

Table 3
Comparison of MLCA Estimates with Prior Published Estimates

Classification		MLCA	Chua & Fuller (1982 data)	Poterba & Summers (1981 data)	CPS Reconciled Reinterview (1977-1982)
True	Observed				
Employed	Emp	98.77 (1993) 98.73 (1995) 98.73 (1996)	98.66 (month 1) 98.65 (month 2)	97.74	98.78
	Unemp	0.34 (1993) 0.49 (1995) 0.37 (1996)	0.32 (month 1) 0.34 (month 2)	0.54	0.19
	NLF	0.89 (1993) 0.78 (1995) 0.79 (1996)	1.02 (month 1) 1.01 (month 2)	1.72	1.03
Unemp	Emp	7.06 (1993) 7.86 (1995) 8.57 (1996)	3.52 (month 1) 3.51 (month 2)	3.78	1.91
	Unemp	81.81 (1993) 76.09 (1995) 74.42 (1996)	88.27 (month 1) 88.23 (month 2)	84.76	88.57
	NLF	11.13 (1993) 16.04 (1995) 17.00 (1996)	8.21 (month 1) 8.16 (month 2)	11.46	9.53
NLF	Emp	1.41 (1993) 1.11 (1995) 1.13 (1996)	1.60 (month 1) 1.61 (month 2)	1.16	0.5
	Unemp	0.75 (1993) 0.69 (1995) 0.87 (1996)	1.19 (month 1) 1.24 (month 2)	0.64	0.29
	NLF	97.84 (1993) 98.20 (1995) 98.00 (1996)	97.21 (month 1) 97.15 (month 2)	98.2	99.21

The table indicates that misclassification of the unemployed as NLF is a bigger problem than misclassification as Employed. Averaging over all three years, approximately two thirds of the error in classifying the unemployed is misclassification as NLF. But the rates of both types of error are high.

Next, we compare our estimates of the CPS classification probabilities with similar estimates from the literature. In Table 3, the MLCA estimates for each of the three years are compared with estimates from Chua and Fuller (1987), Poterba and Summers (1995), and the CPS reconciled reinterview program. Again, the latter three sets of estimates rely on reinterview data while the MLCA estimates are produced directly from the CPS interview data. In general, the relative magnitude of the MLCA estimates across the labor force categories agrees with the previous estimates. The greatest differences occur for the true unemployed population. For this group, the estimates of response accuracy from the literature are three to seven percentage points higher than corresponding MLCA estimates for 1993, which is the time period that most closely corresponds to the comparison estimates.

One explanation for this difference is that the comparison estimates are biased upward as a result of correlations between the errors in interview and reinterview. Another explanation is that the MLCA estimates are biased downward as a result of the failure of the Markov assumption to hold. We suspect that both explanations may be true to some extent. However, the next section provides some evidence that failure of the Markov assumption likely has a small effect on estimates of classification error.

3.6 Robustness of MLCA to Non-Markov Labor Force Transitions

A number of authors have investigated the effects of current and previous employment status on future employment status (see, for example, Akerlof and Main 1980; Heckman and Borjas 1980; Lynch 1989, and Corak 1993). Heckman and Borjas show that examination of this issue is quite difficult due to selection biases, response error, and unobserved heterogeneity. These confounding influences may account for the inconsistent findings in the literature. For example, using data from the CPS, Akerlof and Main (1980) provide evidence that the probability of

future unemployment depends upon the number of previous unemployment spells experienced as well as the duration of those spells. However, in a study of male high school graduates, Heckman and Borjas (1980) found "no evidence that previous occurrences of unemployment or their duration affect future labor market behavior once we control for sample selection bias and heterogeneity bias." The results from the literature are also inconsistent and ambiguous regarding the extent to which the Markov assumption expressed in (2) may be violated for the CPS and other labor market surveys. Nevertheless, in this section, we attempt to provide at least a partial answer to question of how non-Markov labor force transitions affect MLCA estimates of classification error.

To investigate the effect of violations of the Markov assumption in (2) for the present application, we conducted a limited simulation study. To focus the investigation while simplifying the simulation framework, we considered latent structures involving only two classes or states at each time point: unemployed, denoted by X, Y, or $Z = 1$, and other (*i.e.*, employed or not in the labor force), denoted by X, Y, or $Z = 2$ with analogous definitions for the observed states A, B, and C. To create a population for the simulation, the latent probabilities $\pi_x, \pi_{y|x}$, and $\pi_{z|xy}$ and the response probabilities $\pi_{a|x} = \pi_{b|y} = \pi_{c|z}$ were specified to be consistent with the combined 1993, 1995, and 1996 data sets.

We then defined two parameters, λ_1 and λ_2 to be varied in the simulation, where

$$\lambda_1 = \frac{\pi_{z=1|x=2,y=1}}{\pi_{z=1|x=1,y=1}} \tag{17}$$

and

$$\lambda_2 = \frac{\pi_{z=1|x=2,y=2}}{\pi_{z=1|x=1,y=2}}. \tag{18}$$

Thus, λ_1 is the probability of being "unemployed" in March, given "unemployed" in February and "other" in January over the probability of being "unemployed" in March given "unemployed" in the two previous months. Consistent with the findings of Akerlof and Main (1980) who showed that the likelihood of remaining unemployed increases as the number of unemployment spells increases, we assume that $0 \le \lambda_1 \le 1$. Similarly, λ_2 is the probability of being "unemployed" in March, given "other" in the two previous months, over the probability of being "unemployed", given "other" in February and "unemployed" in January. Again, by Akerlof and Main, we assume $0 \le \lambda_2 \le 1$. Note that when $\lambda_1 = \lambda_2 = 1$, unemployment transitions from February to March are Markov.

The simulated data were generated to be completely consistent with a MLCA model having non-stationary transition probabilities when $\lambda_1 = \lambda_2 = 1$. We simulated failure of the Markov assumption by varying λ_1 and λ_2 between 0 and 1. To be consistent with the 1993-1996 data, we fixed the probability of a correct "unemployed" response, $\pi_{a=1|x=1}$, at 0.80 and the probability of a correct "other" response, $\pi_{a=2|x=2}$, at 0.99 in all simulations. In addition, the denominators of λ_1 and λ_2 were fixed to their values as determined from the combined 1993-1996 data while the numerators were computed from (17) and (18) using the values of λ_1 and λ_2 specified in each simulation run.

Table 4 summarizes the results of the simulation for $\lambda_1 = \lambda_2 = \lambda$ where λ is varied from 0.2 to 1.0 in steps of 0.2. Note that for $\lambda_1 = \lambda_2 = 1.0$, which corresponds to a Markov model, the estimated probabilities of correct response are exactly as specified. For smaller values of λ_1 and λ_2, the estimates become negatively biased and are most biased for the lowest value considered, 0.2. Nevertheless, the absolute biases due to non-Markov transitions probabilities are never more than 3 percentage points. The results in Table 4 are consistent with Bushery and Kindelberger (1999), who used a somewhat different approach to illustrate the same robustness property of the MLCA models for CPS data. Both studies suggest that failure of the Markov assumption to hold does not appear to be an important source of bias in estimating CPS classification error probabilities.

Table 4
Estimates of Correct Classification Under Non-Markov Transitions
(Cell entries are percentages)

Pr (Correct)	$\lambda_1 = \lambda_2 = \lambda$				
	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$ (Markov)
Pr ("unemp"| true "unemp") = $\pi_{a=1\|x=1}$	77.6	78.1	78.7	79.3	80
Pr ("other"| true "other") = $\pi_{a=2\|x=2}$	98.6	98.7	98.8	98.9	99

4. COMPARING THE MLCA AND UNRECONCILED REINTERVIEW ESTIMATES

4.1 Hui-Walter Estimation

An alternative set of response probability estimates can be obtained from the CPS reinterview data using a type of latent class model first proposed by Hui and Walter (1980). Using the notation introduced above, let X denote the true labor force classification for some time point and let A and A' denote the interview and reinterview classifications, respectively. Let G denote a grouping variable defined as in (4). Consider the likelihood of the group× interview× reinterview table denoted by GAA'. Denote by $\pi_{gaa'}$ the probability of classifying an individual belonging to group g into cell (a, a') of the table. The model for $\pi_{gaa'}$ proposed by Hui and Walter is

$$\pi_{gaa'} = \sum_x \pi_g \pi_{x|g} \pi_{a|x} \pi_{a'|x}. \tag{19}$$

In this model, the parallel measures assumption for the interview and reinterview responses is relaxed and response probabilities for the two measures, *viz.* $\pi_{a|x}$ and $\pi_{a'|x}$, are estimated separately. The ICE assumption is made as a condition of identifiability. It is further assumed that $\pi_{a|x}$ and $\pi_{a'|x}$ do not depend upon the group variable, G, while the prevalence of employed, unemployed, and NLF, *i.e.* $\pi_{x|g}$, still depends upon G.

Sinclair and Gastwirth's (1996) analysis of CPS labor force classification error used Sex as the grouping variable and our analysis uses this grouping variable as well. Sinclair and Gastwirth confined their analysis to white males and females and two labor force categories: NLF and In the Labor Force. The latter category is the sum of our Employed and Unemployed categories. In our analysis, we consider sample members of all races and analyze the three category labor force classification used in the MLCA. Thus, the H-W analysis estimates 16 parameters for each year, which equals the number of degrees of freedom available from the $G \times A \times A'$ table, leaving no degrees of freedom to test model fit.

The ℓEM software was used to fit the H-W model to the interview and unreconciled reinterview data from three time periods that coincide with the three in our MLCA: pre-1994, 1995, and 1996. We attempted to restrict the analysis to only the first quarter of these time periods. Unfortunately due the small sample sizes, the estimates were quite unstable. Thus, it was necessary to use the reinterview data from all four quarters of these time periods. The pre-1994 data were collected from 1985 through 1988 using the unreconciled reinterview sample.

The results of this comparison of MLCA and H-W estimates are summarized in Table 5. The MLCA estimates are the same as those in the rows of Table 2 labeled "Total." The H-W estimates are the classification probabilities associated with the original interview, *i.e.*, measure A in (19). The table shows the comparison for all three years. Since the largest error rate in the MLCA occurred for the Unemployed, this category is of particular interest in the MLCA/H-W comparison.

Overall, the two sets of estimates show fairly good agreement. The years 1995 and 1996 exhibit no statistically significant differences (at the 5 percent level) between the MLCA and H-W estimates for the unemployed population. The pre-1994 estimates display significant differences; however, they may be explained by the fact that the pre-1994 reinterview data were from 1985 through 1988, rather than 1993. These differences will be explored further in the next section.

4.2 Comparison of Indexes of Inconsistency

As described in section 3.1, we compute estimates of the index of inconsistency for each time period using the MLCA model-based estimates of the response probabilities. Essentially, we estimate the expected interview-reinterview cross-classification table from the MLCA response probability estimates and then apply the formula for the index to this table as though the table were observed. A second expected interview-reinterview classification table be can estimated using the H-W response probability estimates. We then compared these two sets of estimates to the estimate of the index computed directly from the CPS reinterview data using traditional methods (U.S. Bureau of the Census 1985). Agreement of all the three estimates agree supports the validity of the three methods.

Table 5
Comparison of MLCA and H-W Model Estimates of CPS Response Probabilities by Year
(Standards Errors are in Parentheses)

Classification		1993		1995		1996	
True	Observed	H-W	MLCA	H-W	MLCA	H-W	MLCA
Emp	Emp	99.3	98.8	99.5	98.7	99.6	98.8
		(0.3)	(0.1)	(0.7)	(0.1)	(0.1)	(0.1)
	Unemp	0.0	0.3	0.0	0.5	0.4	0.4
		(0.0)	(0.1)	(n/a)	(0.1)	(0.1)	(0.1)
	NLF	0.7	0.9	0.5	0.8	0.0	0.8
		(0.3)	(0.1)	(0.7)	(0.1)	(n/a)	(0.1)
Unemp	Emp	11.1	7.1	11.5	7.9	4.6	8.6
		(1.0)	(0.7)	(2.3)	(0.9)	(15.2)	(1.0)
	Unemp	74.3	81.8	67.9	76.1	67.6	74.4
		(2.7)	(1.1)	(6.1)	(1.3)	(11.1)	(1.4)
	NLF	14.7	11.1	20.6	16.0	27.9	17.0
		(2.9)	(0.9)	(6.5)	(1.2)	(5.3)	(1.2)
NLF	Emp	2.0	1.4	2.5	1.1	2.6	1.1
		(0.5)	(0.1)	(1.5)	(0.1)	(1.5)	(0.1)
	Unemp	1.2	0.8	0.5	0.7	0.0	0.9
		(0.3)	(0.1)	(0.6)	(0.1)	(n/a)	(0.1)
	NLF	96.8	97.8	97.0	98.2	97.4	98.0
		(0.6)	(0.1)	(1.6)	(0.1)	(1.1)	(0.1)

Table 6 shows the three methods estimates the index of inconsistency for all three time periods. As before, the Unemployed category is of particular interest because of its large error rate. Standard errors are not available for the MLCA or the H-W estimates of the index so formal tests of hypothesis are not possible. However, standard errors for the traditional estimates are provided which can be used as rough approximations of the standard errors for the H-W estimates.

Overall, both the general patterns of the MLCA estimates and the magnitudes of the MLCA estimates generally agree quite well with the H-W and traditional estimates for all three years. However, for the NLF category in 1995 and 1996, the traditional estimates of I are somewhat larger than either of the latent class model estimates. Further analysis suggests that this difference is due to a bias in the traditional estimation approach resulting from the failure of the parallel measures assumption.

U.S. Bureau of the Census (1985) shows that if the interview and reinterview processes have different reliabilities, then the traditional estimate of the index will be biased. For example, if the reliability of the reinterview

data is lower than the reliability of the interview data, the traditional test-retest reliability estimator will understate the actual reliability of the CPS data; *i.e.*, the CPS index of inconsistency will be too large.

Table 6
Comparison of MLCA, H-W, and Traditional Estimates of the Index of Inconsistency by Year and Labor Force Classification

Method of Estimation	Labor Force Classification			Aggregate Index
	Employed	Unemployed	Not in Labor Force	
1993				
Traditional estimation	8.16 (0.24)	33.49 (1.16)	9.96 (0.27)	11.05 (0.26)
H-W	7.37	34.93	10.07	10.78
MLCA	6.35	28.04	7.63	8.73
1995				
Traditional estimation	6.69 (0.44)	36.28 (2.85)	10.80 (0.56)	10.42 (0.53)
H-W	6.82	37	8.98	9.7
MLCA	6.06	36.19	7.2	8.72
1996				
Traditional estimation	5.93 (0.39)	35.97 (2.68)	11.95 (0.56)	10.61 (0.51)
H-W	5.67	39.46	7.55	8.56
MLCA	5.99	37.39	7.76	9.06

The CPS interview and reinterview will have different reliabilities if the error distributions for the two interviews are not equal. A test of this is possible by comparing the fit of a H-W type model with and without the restriction $\pi_{a|x} = \pi_{a'|x}$. The assumption of equal reliability is rejected if the difference between the likelihood ratio chi-squares for the two models exceeds a chi-square with 6 degrees of freedom. This test was rejected for 1995 and 1996 at the 10 percent level of significance. Thus, it appears that the difference in the NLF estimates for 1995 and 1996 may be due, in part, to bias in the traditional estimates of *I*.

Note further that the H-W and MLCA indexes agree quite well for 1995 and 1996, although they differ somewhat in 1993. However, as noted in the discussion of Table 5, the comparisons between the MLCA and H-W estimates for this year are confounded by the different time periods used to construct the pre-1994 interview-reinterview data set. This could account for at least some of the discrepancy between the estimates for this year.

5. SUMMARY AND CONCLUSIONS

The primary goal of this research was to investigate the validity of MLCA estimates of CPS labor force classification error and to determine the efficacy of MLCA as an alternative to traditional methods for evaluating CPS data quality. We analyzed interview data from the CPS for the first quarter of three years – 1993, 1995, and 1996 – and conducted an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provided another approach for estimating CPS classification error that, when compared with the MLCA estimates, helped to address the question of the validity of the MLCA approach.

Five dimensions of MLCA validity were addressed as follows:

1. **Model diagnostics.** We investigated a wide range of MLCA models with grouping variables defined by age, race, sex, education, mode of interview, and proxy/self response. The most parsimonious and best fitting model for all three years included one grouping variable defined by the proxy/self variable with four categories: all three waves conducted by self response, only two waves conducted by some self response, only two waves conducted by proxy response, and all three waves conducted by proxy response. For this class of models, the best model was Model 4 (see Table 1) which specified non-homogeneous and non-stationary transition probabilities and non-homogeneous response probabilities. This model provided an adequate fit to the data for all three years.

2. **Model Goodness of Fit Across Years of CPS.** Another indicator of model validity is its fit across independent samples of the same population. Assuming that labor force dynamics and the response probability structure for the CPS is stable across the span of four years, the same general model should fit all three years adequately. Model 4 displays multi-year goodness of fit (see Table 1). In addition, other grouping variables were tested in the study, yet the proxy/self variable model emerged as the best variable for all three years.

3. **Agreement Between the Model and Test-Retest Estimates of Response Probabilities.** Using the unreconciled interview-reinterview data from the CPS for the time periods pre-1994, 1995, and 1996, we applied the H-W method to estimate the response probabilities and compared these with the MLCA estimates. There was good agreement for 1995 and 1996, the two years for which the time periods for the reinterview data and the CPS data were closely matched (see Table 5). For 1993, we observed small but significant differences between MLCA estimates and the corresponding H-W estimates. These differences might be explained by differences in the time periods, since the reinterview data predated the CPS interview data by some years.

4. **Agreement Between the Model and Test-Retest Estimates of Inconsistency.** We compared MLCA model-based estimates of the index of inconsistency with the corresponding direct estimates from the CPS

reinterview program. The two sets of estimates agree fairly well for all three years, with the exception of the NLF category (see Table 6). For 1995 and 1996, the differences can be partly explained by the bias in the traditional estimator resulting from the failure of the parallel measures assumption. The H-W method, which does not require the assumption of parallel measures, produces estimates of the index that agree well with MLCA estimates for 1995 and 1996. For 1993, the difference between the MLCA and H-W estimates may be due to the difference in the time periods for the reinterview and the CPS data sets.

5. **Plausibility of the Patterns of Classification Error.** The MLCA estimates of misclassification probabilities appear to be plausible. The estimates across proxy/self groups were consistent with prior expectations that lower error rates should be observed for self respondents than for proxy respondents. In addition, the largest error rates were observed for the unemployed population and the magnitudes of these estimates were consistent with those of previous studies – for *e.g.*, Fuller and Chua 1985; Abowd and Zellner 1985; Porterba and Summers 1986; and Sinclair and Gastwirth 1996 (see Table 3).

In summary, we found no evidence from these analyses to question the validity of the MLCA approach. The method performed well in all five validity tests. We therefore recommend that the MLCA method be considered as an alternative method for evaluating the accuracy of the CPS labor force estimates. The strong agreement between the MLCA and H-W estimates supports the validity of the H-W method as well. We recommend that both methodologies be considered in future studies of CPS data quality.

Although the MLCA approach performed well in our tests, we recommend caution in applying the methodology in other settings. In our analysis, reinterview data provided a means for assessing the validity of the MLCA estimates. However, reinterview data are typically not available in panel surveys and, consequently, analysts may only be able to apply criteria (1), (2), and (5) above to check model validity. The Markov assumption is key to the MLCA approach. Some panel data may seriously violate this assumption. Fortunately, failure of Markov assumption appears not to be an important factor in the validity of MLCA estimates of CPS labor force classification error (*cf.* Table 4).

REFERENCES

ABOWD, J., and ZELLNER, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

AKERLOF, G.A., and MAIN, G.M. (1980). Unemployment spells and unemployment experience. *The American Economic Review*, 70, 3, 885-893.

BAILAR, B. A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BIEMER, P., BUSHERY, J., and FLANAGAN, P. (1997). An Application of Latent Markov Models to the CPS. Internal U.S. Bureau of the Census Technical Report.

BIEMER, P., and FORSMAN, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

BOHRNSTEDT, G.W. (1983). Measurement. *Handbook of Survey Research*, (P.H. Rossi, R.A. Wright, and A.B. Anderson, Eds.). New York: Academic Press.

BUSHERY, J., and KINDELBERGER, K. (1999). Simulation Examples for MLC Analysis. Internal U.S. Bureau of the Census Memorandum, Washington, DC, 70-122.

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, 397, 46-51.

CLOGG, C., and ELIASON, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 210, 37-46.

CORAK, M. (1993). Is unemployment insurance addictive? Evidence from the benefit durations of repeat users. *Industrial and Labor Relations Review*, 47, 1, 62-72.

FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. *Measurement Errors in Surveys*, (P.P. Biemer, *et al.*, Eds.). New York: John Wiley & Sons. 279-302.

FULLER, W., and CHUA, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*. Washington, D.C., U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 65-77.

HECKMAN, J.J., and BORJAS, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions, and answers from a continuous time model of heterogeneous and state dependence. *Economica*, 47, 247-283.

HESS, J., SINGER, E., and BUSHERY, J. (2000). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, II, 4, 346-360.

HUI, S.L., and WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

KLEINBAUM, D.G., KUPPER, L.L., and MULLER, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. Boston: PWS-KENT Publishing Co.

LIU, T.H., and DAYTON, C.M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22, 249 - 264.

LYNCH, L.M. (1989). The youth labor market in the eighties: determinants of re-employment probabilities for young men and women. *The Review of Economics and Statistics*, 37-45.

MEYERS, B. D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 3, 385-390.

MOORE, J.C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics*, 4, 2, 155-122.

O'MUIRCHEARTAIGH, C. (1991). Simple response Variance: Estimation and Determinants. *Measurement Errors in Surveys*, (P. Biemer *et al.*, Eds.). New York: John Wiley & Sons, 551-574.

POTERBA, J., and SUMMERS, L. (1986). Reporting errors and labor market dynamics. *Econometrics*, 54, 6, 1319-1338.

POTERBA, J., and SUMMERS, L. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.

POULSEN, C.S. (1982). Latent Structure Analysis with Choice Modeling Applications. Doctoral dissertation, Wharton School, University of Pennsylvania.

ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.

SCHREINER, I. (1980). Reinterview Results from the CPS Independent Reconciliation Experiment (Second Quarter 1978 through Third Quarter 1979). Internal U.S. Bureau of the Census Report.

SHOCKEY, J. (1988). Adjusting for response error in panel surveys, a latent class approach. *Sociological Methods and Research*, 17, 1, 65-92.

SINCLAIR, M., and GASTWIRTH, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.

SINCLAIR, M., and GASTWIRTH, J. (1998). Estimates of the errors in classification in the labour force survey and their effects on the reported unemployment rate. *Survey Methodology*, 24, 2, 157-169.

SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, 90, 430, 478-488.

U.S. BUREAU OF THE CENSUS (1985). Evaluating Censuses of Population and Housing, STD-ISP-TR-5. Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (2000). Current Population Survey: Design and Methodology. U.S. Bureau of the Census Technical Paper 63, Washington, D.C.: Government Printing Office.

VAN DE POL, F., and DE LEEUW, J. (1986). A latent markov model to correct for measurement error. *Sociological Methods and Research*, 15, 1-2, 118-141.

VAN DE POL, F., and LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. *Survey Measurement and Process Quality*, (L. Lyberg, *et al.*, Eds.). New York: John Wiley & Sons.

VERMUNT, J. (1997). ℓEM: A General Program for the Analysis of Categorical Data. Tilburg University.

WIGGINS, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Amsterdam: Elsevier S.P.C.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 153-162
Statistics Canada

Estimation and Replicate Variance Estimation of Median Sales Prices of Sold Houses

KATHERINE J. THOMPSON and RICHARD S. SIGMAN[1]

ABSTRACT

The U.S. Census Bureau publishes estimates of medians for several characteristics of new houses, with a key estimate being sales price of sold houses. These estimates are calculated from data acquired from interviews of home builders by the Survey of Construction (SOC). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample replication (MHS) method of variance estimation. The literature supports applying the MHS method to replicate sample medians to estimate the sampling variance of a median. There are several computational advantages, however, to using grouped data to estimate medians, with linear interpolation being used within the grouped-data interval containing the median. Using survey data and simulated finite populations, we compared the effects of no grouping (*i.e.*, the sample median), grouping with fixed-size intervals, and grouping with data-dependent-sized intervals on medians and associated MHS variance estimates. We examined the mean squared errors and mean absolute errors of the median estimates and the relative bias and stability of the variance estimates and the coverage of the associated confidence intervals. We found that the data-dependent-sized intervals yielded variance estimates with the smallest bias, the best stability, and the best confidence intervals.

KEY WORDS: Median; Modified half-sample replication; Survey of Construction.

1. INTRODUCTION

The U.S. Census Bureau publishes estimates of medians for several characteristics of new houses, with a key estimate being sales price of sold houses. These estimates are calculated from data acquired from interviews of home builders by the Survey of Construction (SOC). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample (MHS) replication method (balanced repeated replication with replicate weights of 1.5 and 0.5) for reasons outlined in section 3.B. In the near future, the SOC will move its current estimation and variance estimation systems to the Census Bureau's re-engineered post-data-collection system, the Standardized Economic Processing System (StEPS). When this occurs, SOC will change from its current non-replicate variance estimation procedure to the MHS replication variance estimation procedure (Thompson 1998). Because the SOC variance estimation methodology is changing, we decided to revisit the median-estimation methodology for continuous data. Our goal was to find a median-estimation method with good estimation and variance estimation properties, given the MHS replication.

We considered two methods of median-estimation. The first method uses the sample weights to estimate medians via empirical cumulative-distribution functions. The second method uses linear interpolation of grouped continuous data to approximate the median. The latter method is implemented in VPLX (Variances from ComPLeX Survey, Fay 1995), the replicate variance estimation software package developed at the Census Bureau.

Direct calculation of sample medians can be computationally intensive because it requires separate sorts for each value of a given classification variable. An alternative estimation method is to group the continuous data into discrete intervals (called bins) and use linear interpolation over the interval containing the median. Provided that the data are approximately uniformly distributed over the interval containing the median, interpolation yields a good approximation while being considerably less computer resource-demanding. However, optimal bin widths and locations may differ by domain and may change over time as the sample distributions change.

In this paper, we compare six methods of median-estimation, given MHS replication: the sample median and five variations using linear interpolation. Section 2 provides a brief overview of the SOC design. Section 3 presents general methodology. Section 4 describes the empirical results from four months of SOC data that motivated the simulation study presented in section 5. Section 6 provides our conclusions and recommendations.

2. SOC SAMPLE DESIGN

The SOC universe contains two sub-populations: local areas that require building permits and local areas that do not. The SOC sample-units selected from the first sub-population comprise the Survey of the Use of Permits (SUP), and those selected from the second sub-population, the Nonpermit Survey (NP). The SUP sample comprises the majority of the SOC estimate. The two samples are multi-stage probability samples stratified by variables with high expected correlation with the survey's key statistics: housing starts, completions, and sales.

[1] Katherine J. Thompson and Richard S. Sigman, Economic Statistical Methods and Programming Division, U.S. Census Bureau, Washington DC, 20233, U.S.A.

The first stage of the SUP and NP sample selection is a subsample of 1980 design Current Population Survey (CPS) Primary Sampling Units (PSUs), which are contiguous areas of land with well-defined boundaries. Thus, both surveys are conducted in the same PSUs but are otherwise independent samples. The PSUs were stratified within region by weighted 1980 population 16 years and older, weighted 1982 residential permit activity, and percent of housing in nonpermit areas. When possible, strata consisted of PSUs from the same state with the same metropolitan status. One PSU per stratum was selected. Self-representing (SR) PSUs were included in the sample with certainty (the stratum consists of one PSU). Nonself-representing (NSR) PSUs were selected with probability proportional to size (PPS) from strata containing more than one PSU.

The second stage of SUP sample selection is a stratified systematic sample of permit-issuing places within sample PSUs (selected once a decade). These places were stratified by a weighted average of the ratio of permit-issuing activity in year $\underline{i}$ to the total US permit activity in year $\underline{i}$ ($\underline{i}$ = 78, 81, 82). In many cases, only one second stage unit was selected. The third stage of SUP sample selection is performed monthly: each month, Field Representatives (FRs) select a systematic sample of building permits from the permit offices in each sampled permit-issuing place. One-to-four-unit building permits are selected systematically in such a way that an overall one-in-forty sample is achieved; five-or-more-unit building permits are included with certainty. The third-stage samples are independent by month; the first and second stages are not.

The second stage of NP sample selection is a stratified systematic sample of small land areas (1980 Census Enumeration Districts, or EDs), stratified by 1980 Census population size. For the third stage of NP sample selection, field representatives completely canvass all of the roads in the sampled EDs (called segments). To reduce canvassing, a few of the larger EDs were subsegmented and one subsegment selected, or large EDs were 1-in-2 subsampled. Currently, there are a total of seventy-one active nonpermit segments. All new housing units are included in the NP sample with certainty.

Median estimates are derived from the pooled SUP and NP samples and are calculated using a post-stratified weight for the SUP portion and an unbiased weight for the NP portion.

3. METHODOLOGY

A. Median-Estimation Procedures

1. Sample Median

One procedure for estimating the median of a population is to calculate the sample median from ungrouped data, using the sample weight to locate the median. This approach is recommended in Kovar, Rao and Wu (1988) and Rao and Shao (1996). The procedure uses the following steps:

- sort the sample observations in ascending order;
- accumulate the sum of the associated survey weights;
- select the first observation for which the associated sum of the weights exceeds fifty percent of the total weight.

2. Linear Interpolation

Another approach for estimating the median of a population is to group the sample data and interpolate for the sample median. Woodruff (1952) provides the following formula for linear interpolation of a sample median:

$$\hat{M} = F^{-1}(\frac{1}{2}\hat{N}) \approx ll + \left(\frac{\frac{1}{2}\hat{N} - \text{cf}}{f_i} \right) * (i) \qquad (3.1)$$

where

F = the cumulative frequency of the characteristic using sample weights

ll = lower limit of the bin containing the median

$\hat{N}$ = estimated total number of elements in the population

cf = cumulative frequency in all intervals preceding the bin containing the median

f_i = median class frequency (estimated total number of elements in the population of the interval containing the median)

i = width of the bin containing the median

This is the method used by the current SOC production variance estimation system for monthly estimates and is also the linear interpolation method employed by VPLX.

We considered two options for setting the class size (bin widths) for the interpolation. The first option develops bins based on the specific characteristic under consideration using the original data. The second option linearly transforms the data to a standard scale and then uses a standard set of bins for every characteristic. We used the following linear transformation:

$$X_i' = X_i * \frac{1{,}000}{Q_3} \qquad (3.2)$$

where Q_3 is the third quartile of the sample distribution (estimated using the ordered observations and sample weight as outlined in section 2.A.1). The interpolated median of the X' is multiplied by $(Q_3/1{,}000)$ to obtain an estimated median on the original scale [If the distribution contains negative values (*e.g.*, a distribution of net income), then use $X_i'' = (X_i - X_{(1)}) * 1{,}000/Q_3(X_i - X_{(1)})$, where $X_{(1)}$ is

the first order statistic and $Q_3(X_i - X_{(1)})$ is calculated from the distribution of $(X_i - X_{(1)})$. To obtain an estimated median on the original scale, multiply the interpolated median by $(Q_3(X_i - X_{(1)})/1{,}000)$ and add $X_{(1)}$.] This procedure is equivalent to simply dividing the original sample from 0 to Q_3 into $\underline{x}$ bins of equal width and placing the remainder of the data into one bin which, by design, is much larger than the others.

This procedure is designed for symmetric or positively skewed distributions (usually the case with economic data). The data in the last bin is not used to estimate the median because it is greater than Q_3, which is expected to be far from the median. If we based the linear transformation on Q_1 (the first quartile), the bin containing the median might be very close to the lowest bin in the distribution. In this case, the difference in variability between an interpolated median and the sample median would be small.

Using the original data to develop medians has the advantage of producing production-ready estimates and SEs. Determining the appropriate fixed bin width is difficult, however. As the bin widths get small (approach width 1), the variance estimates become more unstable. As the bin widths increase, the bias of the estimate due to interpolation increases. The "optimal" bin size balances variance estimate stability and bias. Unfortunately, the optimal bin width may not remain constant between samples. Often, the distributions change over time, and the bins widths/locations in the sample should reflect this change in scale. Moreover, the optimal bin width may be different for different values of a classification variable: for example, the optimal bin width for the Midwest's sales price is probably different from the optimal bin width for the South's sales price.

The desire to have the width of the bin depend on the sample motivated the linear transformation. The "standard" bin widths used for the transformed data less than Q_3 are not standard on the untransformed scale: the bin width is data-dependent. Using the linearly transformed data requires more bookkeeping in terms of scaling constants but easily allows for changes in the scale and shape of the distribution.

Figures 1 through 4 illustrate the effect of the linear transformation on the bin widths and location for two distributions. Figures 1 and 2 present a distribution that has a large spread of data values, including a few very large observations. Figures 3 and 4 present a distribution consisting of primarily small data values.

Figure 1 presents a histogram of the original distribution for houses sold with conventional financing, with bin width of \$25,000 [Note: the bin size was selected purely for presentation convenience, since this is a long-tailed distribution]. The median of this distribution is \$167,130, and Q_3 is \$225,000. Figure 2 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are equivalent to bins of width \$11,250 on the original scale (($225,000/1,000)*50). Recall that the original-data bin sizes considered are \$1,000 and \$2,000. Thus, the transformed-data bins of width 4 would have a width of \$900 on the original untransformed scale. Notice the large "spike" at the last bin, which contains all of the sample greater than Q_3.

These figures also illustrate the differences in distribution of sample sizes across bins between the two methods. Using fixed bin widths with the original data results in quite variable bin sample sizes (see Figure 1). In contrast, by design the sample sizes within the data-dependent bins are much more uniform for all but the last bin (see Figure 2).

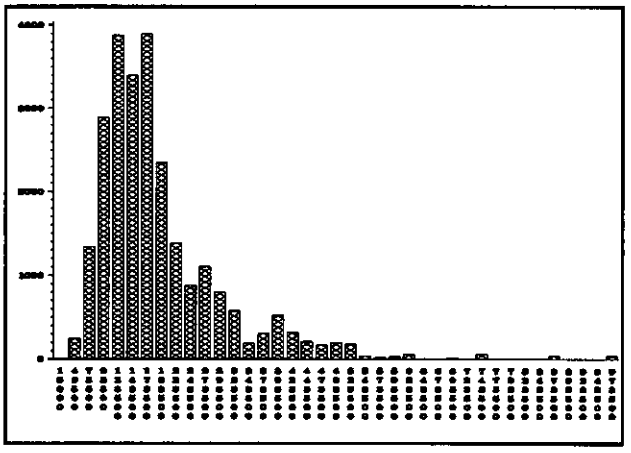


Figure 1: Original Distribution of Sales Price of Houses Sold With Conventional Financing Bin Width = \$25,000

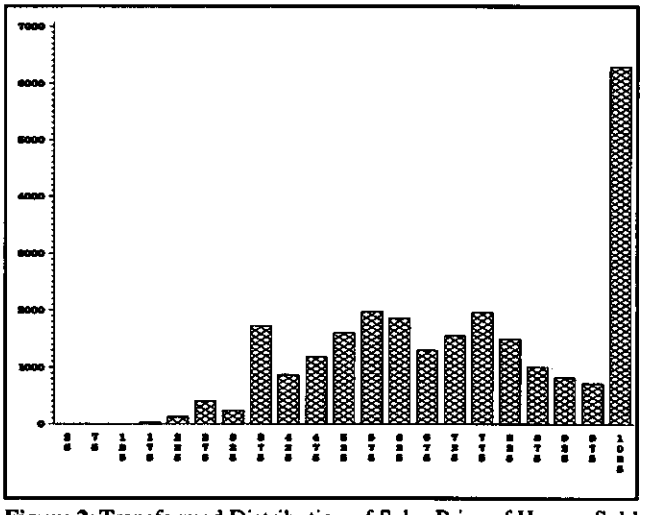


Figure 2: Transformed Distribution of Sales Price of Houses Sold With Conventional Financing Using Bin Width = 50 Bin Width on Untransformed Scale = \$11,250

Figure 3 presents a histogram of the original distribution of houses sold with FHA loans, with bin width of \$4,000 (again, the bin width is chosen for presentation convenience). The median of this distribution is \$108,280, and Q_3 is \$124,990. Figure 4 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are

equivalent to bins of width \$6,250 on the original scale, and the transformed-data bins of width 4 would have approximate width \$500 on the original untransformed scale.

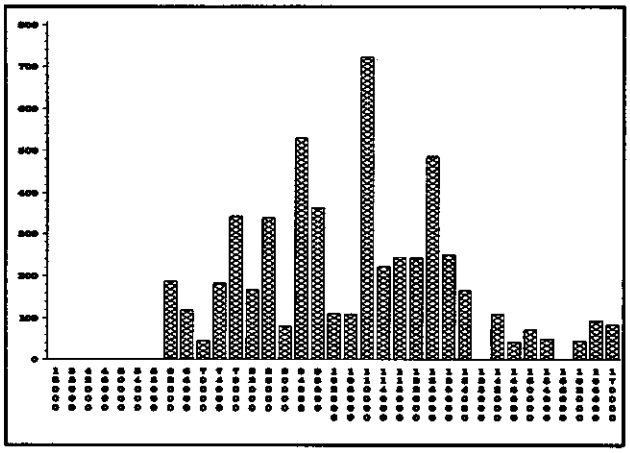


Figure 3: Original Distribution of Sales Price of Houses Sold With FHA Loans Bin Width = \$4,000

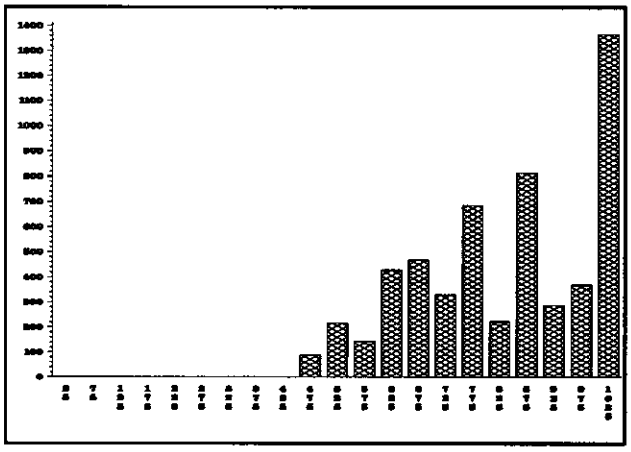


Figure 4: Transformed Distribution of Sales Price of Houses Sold With FHA Loans Using Bin Width = 50
Bin Width on Untransformed Scale = \$6,250

Figures 1 through 4 demonstrate the flexibility of the bins developed for linearly-transformed data. The bin size on the untransformed scale expands or contracts, depending on the spread of the data. Moreover, the data-dependent bin sample sizes are less variable compared to those associated with fixed bins.

To evaluate the first interpolation option (original-data-interpolated medians), we used two different sets of bin widths (classification sizes): bins of size \$2,000 (the same bin width used in the current production variance estimation system) and bins of size \$1,000. [Note: The VPLX variance estimation software would not allow any bin size smaller than 1,000 because the number of classes exceeded the allowable array range.] After examining several months of sales price estimates for the total U.S., we assumed that median sales price would always be larger than \$36,000 and smaller than \$550,000, so the first original-data classification is always (low – 35,999) and the last original-data classification is always (550,000 – high): this yields 257 bins of size \$2,000 or 514 bins of size \$1,000, plus one bin of size \$36,000 and one bin whose width depends on the largest observation in the sample. One obvious problem with the locations of these bins is the potential effect of inflation. It is conceivable that within special financing categories or certain regions, the median sales price for houses sold could approach \$550,000, and the interpolation would fail as a consequence.

To evaluate the second interpolation option (transformed-data-interpolated-medians), we used three different sets of bin widths: bins of size 4, 25, and 50. The bins of size 4 were chosen to be analogous to the bins of size 2,000 in terms of the number of bins. There are 250 bins of size 4 for the transformed data less than Q_3, and one larger bin containing all data greater than Q_3. The selection of widths 25 and 50 was somewhat arbitrary: we chose bin size 50 to get a total of twenty bins for the data less than Q_3; and we chose bin size 25 to examine the effect of doubling the number of bins/halving the width of the bins for data less than Q_3. The transformed-data median is always less than 1,000, so the last transformed-data classification is always (1,000 – high). Thus, by definition the last bin contains up to twenty-five percent of the data and is considerably wider than the other bins.

B. Variance Estimation

We used the Modified Half-Sample (MHS) replication method (Fay 1989 and Judkins 1990) to estimate the variance of a median as supported in the literature (*e.g.*, Rao, Wu, and Yue (1992); Rao and Shao (1996); Kovacevic and Yung (1997) for balanced repeated replication; and Judkins (1990) for MHS replication). MHS replication is a variation of the "traditional" balanced half-sample variance estimation described in Wolter (1985, 110-152). Balanced half-sample replication (BRR) is a variance estimation method designed for a two-PSU per stratum design. With BRR, a half-sample replicate is formed by selecting one unit from each pair and weighting the selected unit by 2 (so that it represents both units). Thus, estimates for every PSU are included in each replicate although half are weighted by zero. Replicates (half-samples) are specified using a Hadamard matrix. See Wolter (1985, 114-115) for a detailed description of the replicate formation procedure using Hadamard matrices. MHS replication uses replicate weights of 1.5 and 0.5 in place of the 2 and 0. The standard error for a median estimate using MHS replication is given by

$$\hat{SE}(\hat{Med}) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R} (\hat{Med}_r - \hat{Med}_0)^2}$$

where the r subscript refers to the replicate r median estimate ($r = 1, 2, ..., R$) and the 0 subscript refers to the full sample the median estimate. This expression contains a four (4) in the numerator because the MSE of the replicate

estimates is too small by a factor of $1/(1-0.5)^2$. See Judkins (1990).

Neither the SUP nor the NP designs are two-sample-unit-per-stratum designs. At the first stage, one PSU per stratum is selected. The second and third stages are systematic samples, and often only one unit per stratum was selected at the second stage. A common approach used to address the one sample-unit per stratum problem is to

- "split" the SR sample-units into two panels per sample-unit using the original sampling methodology;
- form collapsed strata by pairing two (or three) "similar" NSR sample-units; and
- apply the half-sample approach in such a way that the elements contributing to the half samples are panels within sample-units for SR sample-units and are the first stage sample-units (PSUs) within collapsed strata for NSR sample-units.

The current SOC production variance system uses a Keyfitz estimator (a paired difference estimator) for NSR sample and an approximate sampling-formula estimator for SR sample to produce level estimate variances (Luery 1990). Because SOC methodologists had already collapsed NSR strata for their paired difference estimator, a BRR-like application was a logical extension of the pre-existing variance estimation structure. For MHS replication, we sort permits within predetermined sample-unit groups in SR units by geography and authorization date and systematically split the ordered sample into two panels as suggested in Wolter (1985, 131). Although this is essentially the only approach available for the SOC design, this method may not provide the correct variance estimates since units in both panels are correlated (in the original half-sample method, the two PSUs in the stratum are assumed independent). For more details on the replicate assignments, see Thompson (1998).

The SOC production system uses the Woodruff method (Woodruff 1952) to estimate the standard error of a median. The Woodruff method uses the estimated SE of a proportion $\underline{p}$ ($\underline{p} = 0.50$ for median-estimation) and projects the interval ($\underline{p} \pm SE(\underline{p})$) through the cumulative frequency distribution to obtain the lower limit of a 62.86 percent confidence interval for the median (the $SE(\underline{p})$ can be estimated using replicate methods). The SE of the median is then estimated by subtraction. This methodology has had mixed success in the past according to SOC survey analysts.

4. EMPIRICAL DATA RESULTS

Initially, we used four months of SOC sample data to examine the variances of the median-estimation methods for sales price of sold houses: March 1997, May 1997, June 1997, and July 1997. We produced medians by region and by type of financing. We used the same weight used by the SOC production estimation and variance systems (post-stratified for SUP sample and unbiased for NP sample), pooling both surveys' data to obtain medians. Each set of variance estimates was produced using 200 replicates.

We found that the six median-estimation methods produced very similar estimates, but yielded three distinct sets of SEs: one set for the sample median, one set for the original-data-interpolated medians (fixed bin width), and one set for the transformed-data-interpolated medians (data-dependent bin width). There was no clear relationship between bin width and SE estimates within the two sets of interpolated medians. Indeed, within type of data (original or transformed), the SEs were all very close. Clearly, there was a linear transformation and an interpolation effect. None of the median-estimation methods yielded standard errors resembling the published standard errors, so there was no available argument for publication consistency.

Moreover, there is some evidence that the Woodruff method publication SEs are underestimates or are at least inappropriate for the sample design used. Kovar, Rao, and Wu (1988) compared Woodruff SEs and BRR standard errors and found that the two methods had similar properties except for the case of stratified samples, where the strata are based on highly correlated separate variables (such as the SOC design). In this case, the Woodruff SE is often too small, and they concluded that "the BRR... methods (sic) are more robust to different population structures, since the error is extracted directly from the replicates." When the production system Woodruff SEs used the directly-calculated $SE(\underline{p})$, the Woodruff SEs were generally smaller than the replicate SEs.

The empirical results left us in a quandary. We had three distinct sets of variance estimates, and no "gold standard" against which to measure them. Because our empirical results were inconclusive, we conducted a Monte Carlo simulation study to evaluate the properties of the MHS variance estimates produced from the different median estimators.

5. SIMULATION STUDY COMPARISON

A. Procedure for Simulation Study

We created four finite artificial populations based on a data analysis of four SOC sample populations: one type-of-financing population (Conventional Financing) and three regional populations (Midwest (Region 2), South (Region 3), and West (Region 4)). These populations represented a variety of the types of SOC populations from which estimates are produced. Note that the SOC type-of-financing population is not independent of the SOC-region populations.

To approximate the finite population of sales price for houses sold, we generated w_i records for each sample-unit $\underline{i}$, where w_i is the sample weight associated with unit $\underline{i}$. The distributions of sales price for single-unit sold houses could

be approximated by lognormal distributions. The lognormal distribution has the probability density function

$$f(y) = \frac{1}{y-\theta}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{(\log(y-\theta)-\zeta)}{\sigma}\right)^2\right)$$

$$\text{for } \theta < y < \infty$$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter.

From our models, we generated four simulated finite bivariate populations with expected correlation $\rho = 0.6$ using the method outlined in Naylor, Balintfy, Burdick and Chu (1968, 99). The first of the two variables in each population represented sales price of sold houses and was obtained by generating a random normal variable with mean ζ and variance σ^2 using the parameters determined above, then exponentiating and shifting by the appropriate location parameters (θ). The second variable was used to form strata and first stage clusters. This variable had a marginal standard normal distribution and was obtained by independently generating a second standard random normal value, multiplying it by 0.8, and adding this term to 0.6 × the standard normal random variable used to generate the sales price variable. Percentiles, sample skewness, and sample kurtosis of each simulated population's sales price variable were very close to the corresponding statistics in the original population, especially when outliers were deleted using the resistant outer fences rule described in Hoaglin and Iglewicz (1987). Each population's size was the $\hat{N}$ estimated from the sample populations. Model parameters, sample correlations (between simulated sales price and stratifying variable), population size (N), and sample sizes (n) are reported in Table 1.

After generating the finite populations, we formed 50 equal sized strata in each population, then selected two sets of samples for two different survey designs:

- The first design is patterned after the SUP sample of permits for four-or-less-housing units in SR permit offices in SR PSUs (approximately 28% of the SOC sample). In this study, we selected 5,000 stratified without-replacement random samples from each simulated population using the same sampling rate in each stratum. To perform MHS replication, we sorted the sample within each stratum by stratifying variable and then systematically split the sample into two panels.
- The second design is patterned after the SUP sample of permits for four-or-less-housing units in NSR permit offices in SR PSUs and in SR permit offices in NSR PSUs (approximately 40% of the SOC sample). In this study, we selected 5,000 two-stage samples from each simulated population. The first stage is stratified without-replacement random sample of two PSUs per stratum ($N_h = 5$). The second stage is a systematic sample of units within PSUs. Because all PSUs are the same size, this study does not take the SOC PPS sampling into account and does not include the collapsing of first-stage units. The MHS replication uses the first-stage sample units (PSUs) within the same strata. The replicate weights do not account for large sampling fractions at the first stage of selection as recommended in Wolter (1985, 122), so all of the variance estimates are probably upwardly biased.

We did not attempt to simulate the SUP sample of permits for four-or-less-housing units in NSR PSUs and NSR permit offices (a three-stage sample, approximately 25% of the SOC sample); the SUP sample of permits for five-or-more housing units (approximately 2% of the SOC sample); or the NP sample of EDs (approximately 5% of the SOC sample). The three-stage sample, although non-negligible in SOC, is rarely used by other surveys at the Census Bureau, and the other two sectors of the SOC design do not contribute enough to the estimates to warrant a separate investigation.

To examine the precision of each median-estimation procedure over repeated samples, we estimated empirical Mean Squared Errors (MSE) and Mean Absolute Errors (MAE) from the 5,000 samples for:

SM: the sample median of each half-sample

IO2000: interpolated medians using original data, bins of size 2,000 (fixed bin width)

IO1000: interpolated medians using original data, bins of size 1,000 (fixed bin width)

IT4: interpolated medians using linearly transformed data, bins of size 4 (data dependent bin width)

IT25: interpolated medians using linearly transformed data, bins of size 25 (data dependent bin width)

IT50: interpolated medians using linearly transformed data, bins of size 50 (data dependent bin width)

Table 1
Characteristics of Simulated Populations and Sample Sizes of Stratified Samples

		Sales Price Parameters			Correlation (Stratifier, Sales Price)	Population Size	Sample Size
Population	Distribution	θ	σ	ζ	ρ	N	n
Conventional Financing	lognormal	27,578	0.4895	11.84	0.57030	25,150	500
Midwest	lognormal	31,801	0.5957	11.69	0.55835	6,500	150
South	lognormal	29,414	0.5549	11.55	0.55929	14,550	300
West	lognormal	53,781	0.5822	11.59	0.55525	11,550	250

Table 2
Median, Third Quartile, and Bin Widths on Original Scale for Transformed Simulated Data

Population	Median	Q_3	Bin Width		
			4	25	50
Conventional Financing	167,173	222,263	889	5,557	11,113
Midwest (Region 2)	151,312	210,647	843	5,266	10,532
South (Region 3)	133,745	180,868	723	4,522	9,043
West (Region 4)	162,130	214,320	857	5,358	10,716

The linear transformation was performed once for procedures IT4, IT25, and IT50. The original data were transformed using the full sample Q_3, and these transformed data were assigned to the half-samples (including replicate 0, the full sample). Table 2 provides the median and third quartile of each finite population, along with the bin widths on the original scale for the transformed data.

We calculated $M(\zeta_i)$, the empirical MSE of median-estimation procedure i as

$$M(\zeta_i) = \frac{\sum_r (\zeta_{ri} - \bar{\zeta}_i)^2}{5{,}000} + (\bar{\zeta}_i - \zeta_p)^2 = \hat{\sigma}^2(\zeta_i) + \hat{bias}^2(\zeta_i) \quad (5.1)$$

where ζ_{ri} is the estimated median for sample r and estimator i, $\bar{\zeta}_i$ is the average of the ζ_{ri}, and ζ_p is the population median. This is the empirical MSE described in Judkins (1990).

We calculated the Mean Absolute Error (MAE) of each median-estimation procedure i as

$$MAE(\zeta_i) = \frac{\sum_r |\zeta_{ri} - \zeta_p|}{5{,}000} \quad (5.2)$$

as defined in DeGroot (1986, 209-211).

To compare the variance estimation properties of the different median-estimation methods, we calculated an MHS variance estimate (v_{ij}) corresponding to each median-estimation procedure i from 1,000 of the 5,000 samples. These variance estimates were compared in terms of

Relative bias $(\sum_{j=1} v_{ij}/1{,}000)/M(\zeta_i) - 1$

Relative stability $[(\sum_{j=1} (v_{ij} - M(\zeta_i))^2/1{,}000)]^{1/2}/M(\zeta_i)$

Error Rate Number of samples where ($\zeta_p < \theta_{Li}$ or $\zeta_p > \theta_{Ui}$)/1,000 where

θ_{Li} is the lower end of a 90% confidence interval, and

θ_{Ui} is the upper end of a 90% confidence interval

These criteria are used in Kovar, Rao, and Wu (1988) and in Rao and Shao (1996). The relative bias is a measure of the bias of the variance estimate as a proportion of the true MSE. The stability is a measure of the variance of the variance estimates; it approximates a c.v. of the variance estimate v_i. Note that the relative stability is not the relative MSE defined in Wolter (1985, 297) which uses the squared-MSE in the denominator. With an "optimal" variance estimator, both the relative bias and relative stability will be near zero, and the error rate will be ten percent.

B. Results

1. Comparison of Median-estimation Procedures

   Table 3 presents the empirical root MSE, standard error, the bias, and the MAE for each median-estimation procedure from both simulation studies. Each of these statistics was calculated from 5,000 samples.

   These results reinforced our suspicions from the empirical data analysis described earlier. At least for sales price, all six median-estimation procedures perform approximately equally well, with approximately equal root-MSEs and MAEs between procedures in each population.

2. Comparison of MHS Replication Variance Estimation Properties of Median-Estimation Procedures

When we examined the variance estimation properties for each procedure, the results were quite different. As with our empirical data analysis, we had three very distinctive sets of results. Table 4 summarizes the three different comparison measures for the variance estimates in the four populations. The numerators for the relative bias and stability and the coverage rates are based on 1,000 samples. The denominator for the relative bias and stability ("truth") are based on 5,000 samples. An asterisk (*) in the last column of Table 4 indicates that the error rate is significantly different from the nominal error rate of 0.10 using the normal approximation to the binomial distribution at the 90% confidence level.

Table 3
Empirical Root MSE, Standard Error, Bias, and MAE for Median-Estimation Procedures

Population	Median-Estimation Procedure	Unclustered Single-Stage Sample				Clustered Two-Stage Sample			
		Root MSE	SE	Bias	MAE	Root MSE	SE	Bias	MAE
Conventional Financing	SM	3,345	3,345	-12	2,671	3,389	3,374	324	2,733
	IO2000	3,320	3,316	161	2,698	3,346	3,341	189	2,685
	IO1000	3,387	3,368	-354	2,642	3,431	3,420	-278	2,774
	IT4	3,351	3,340	273	2,673	3,378	3,364	311	2,719
	IT25	3,304	3,293	276	2,617	3,337	3,321	322	2,664
	IT50	3,282	3,265	329	2,606	3,305	3,283	375	2,636
Region 2 Midwest	SM	6,316	6,287	-598	4,966	6,273	6,228	-753	4,959
	IO2000	6,276	6,275	-127	4,992	6,335	6,207	-1,271	5,029
	IO1000	6,343	6,297	-767	4,939	6,526	6,280	-1,774	5,204
	IT4	6,372	6,363	328	5,004	6,294	6,228	-908	4,979
	IT25	6,273	6,272	127	4,937	6,270	6,154	-1,199	4,971
	IT50	6,220	6,218	160	4,936	6,224	6,114	-1,164	4,966
Region 3 South	SM	3,670	3,658	301	2,931	3,835	3,752	796	3,054
	IO2000	3,708	3,669	539	2,998	3,796	3,739	656	3,011
	IO1000	3,742	3,740	101	2,941	3,809	3,804	212	3,066
	IT4	3,718	3,662	639	2,951	3,814	3,736	766	3,028
	IT25	3,699	3,638	669	2,924	3,793	3,711	787	2,992
	IT50	3,692	3,616	745	2,912	3,778	3,680	856	2,970
Region 4 West	SM	4,385	4,382	-140	3,509	4,394	4,351	616	3,506
	IO2000	4,425	4,421	185	3,578	4,362	4,339	449	3,487
	IO1000	4,477	4,469	-258	3,530	4,411	4,410	-57	3,535
	IT4	4,414	4,403	318	3,514	4,383	4,342	599	3,494
	IT25	4,376	4,364	315	3,460	4,334	4,296	573	3,439
	IT50	4,367	4,350	391	3,455	4,320	4,271	644	3,436

In both studies, the variance estimates of the transformed-data-interpolated medians perform best in terms of relative bias and stability. Specifically,

- The variance estimates of the transformed-data-interpolated medians (IT4, IT25, IT50) have the smallest relative bias. The difference in estimation method is quite pronounced in three of the four populations, where the largest relative bias of the transformed-data-interpolated medians is less than one-half the size of the smallest relative bias of the original-data-interpolated and sample medians. These results are surprisingly strong for the two-stage clustered design, since the variance estimates are expected to be biased upwards (see section 5.A);
- The variance estimates of the interpolated medians had the best stability. The variance estimates of the sample median had the poorest stability in all four populations. This result was expected due to the smoothing effect of interpolation. Again, the transformed-data-interpolated medians generally performed better than the original-data-interpolated medians, although the difference is not as pronounced as in the case of relative bias. Generally, the stability is close with all three bin widths for the transformed-data-interpolated medians.

The results for each median-estimation procedure's confidence interval coverage are not as consistent, varying by design. With the single-stage unclustered design, the confidence intervals constructed from transformed-data-interpolated medians and SEs have the best coverage. In each population, the data-dependent bins (all widths) yield close to nominal or better coverage; in fact, none of these error rates is statistically different from the nominal 10%. The confidence intervals constructed from original-data-interpolated medians and SEs are extremely conservative. Here, the positive bias in the variance estimates makes these intervals unnecessarily wide, thereby reducing the power to make interesting findings. The coverage with the sample median is erratic.

Some of these coverage patterns are repeated in the two-stage clustered design. Again, the coverage with the sample median is erratic, and the coverage rates for the confidence intervals constructed from original-data-interpolated medians are better than nominal (although only significantly better than nominal in two populations). The error rate pattern is quite different for the transformed-data-interpolated medians. In all but the Region 4 population, the coverages rates for the three procedures are worse than nominal. However, with bins of widths 4 and 25, only one error rate is significantly larger than 10%; for bins of width 50, two of these three error rates are significantly larger than 10%. All of the interpolated-data-medians have significantly smaller than nominal error rates in the Region 4 population; consistent with the other population's results, the error rates for the original-data-interpolated medians are the farthest from 10%.

Table 4
Relative Bias and Relative Stability for Variance Estimates, and Error Rates for 90% Confidence Intervals

Population	Median-Estimation Procedure	Unclustered Single Stage Design			Clustered Two-Stage Design		
		Relative Bias	Relative Stability	Error Rate	Relative Bias	Relative Stability	Error Rate
Conventional Financing	SM	0.19	0.69	11.0%	0.11	0.58	15.1%*
	IO2000	0.25	0.35	6.9%*	0.25	0.37	9.0%
	IO1000	0.21	0.32	7.0%*	0.19	0.33	9.3%
	IT4	0.06	0.25	10.0%	0.06	0.27	11.3%
	IT25	0.07	0.25	10.9%	0.06	0.27	11.8%*
	IT50	0.05	0.26	9.5%	0.05	0.28	12.1%*
Region 2 Midwest	SM	0.57	1.24	7.3%*	0.41	1.07	7.9%*
	IO2000	0.33	0.44	6.9%*	0.23	0.35	8.6%
	IO1000	0.30	0.42	7.0%*	0.17	0.30	8.7%
	IT4	0.15	0.41	10.1%	0.14	0.41	11.5%*
	IT25	0.16	0.40	9.8%	0.11	0.37	10.4%
	IT50	0.15	0.42	9.0%	0.11	0.40	10.4%
Region 3 South	SM	0.30	0.88	12.4%*	0.15	0.71	11.1%
	IO2000	0.31	0.42	6.7%*	0.28	0.39	7.5%*
	IO1000	0.29	0.40	6.7%*	0.27	0.38	7.3%*
	IT4	0.04	0.29	11.0%	0.01	0.28	10.8%
	IT25	0.02	0.28	11.0%	-0.01	0.27	11.3%
	IT50	0.01	0.29	11.1%	-0.02	0.28	11.9%*
Region 4 West	SM	0.39	0.98	8.9%	0.25	0.79	8.6%
	IO2000	0.32	0.42	6.2%*	0.31	0.41	5.2%*
	IO1000	0.29	0.39	6.2%*	0.28	0.38	5.2%*
	IT4	0.11	0.32	8.6%	0.10	0.31	7.6%*
	IT25	0.10	0.31	9.4%	0.09	0.30	7.5%*
	IT50	0.08	0.31	9.5%	0.08	0.31	8.3%*

In both studies, the transformed-data-interpolated medians have the best variance estimation properties in terms of relative bias and relative stability by a large margin, regardless of bin width. And, in both studies, the transformed-data-interpolated medians using bins of width 4 or width 25 have excellent confidence interval coverage. Since the transformed-data-interpolated-medians using bins of width 50 or width 25 yielded the "best" estimators in terms of root-MSE and MAE in both studies, using linear interpolation on transformed data with bins of width 25 appears to be the best median-estimation procedure in terms of estimation and variance estimation properties.

6. CONCLUSION

We explored the effect of using variations of two different methods of estimating the median sales price of sold houses: direct estimation versus linear interpolation. Linear interpolation requires classifying continuous data into bins of standard width. This width can be arbitrary, can differ greatly by domain, and may change as the sample distribution changes over time. The linear transformation based on the third quartile appeared to correct this problem. With the transformed data, the bins' widths and locations in the distribution change depending on the data.

Our empirical results indicated that the choice of method has a pronounced impact on the variance estimates given MHS replication. Our simulation study examined the properties of the different median-estimation procedures on the MHS replicate variance estimates. In all four simulated populations, the transformed-data-interpolated medians (data dependent bin widths) performed the best, usually by a wide margin. Most critically, this method greatly reduces the overestimation of the variance. Using bins of width 25 on the transformed scale (41 bins total) yielded the best median sales price estimates and variance estimates, given MHS replication and is our recommended method for the Survey of Construction.

The recommended method has several advantages. First, it is adaptive. It works well for a variety of distributions, because the bin widths themselves depend on the distribution at hand. Second, it saves computing resources by avoiding sorting half-samples. Third, the data-dependent-intervals can be easily incorporated into generalized survey-processing software. Finally, it gives better estimates and

MHS replicate variance estimates (at least for sales price of sold houses). We expect that these results are generalizable for other continuous distributions as well, although obviously this conjecture should be tested on other data sets. Other areas for future research include examining the relationship between sample size and precision of the median estimates, examining alternative bin sizes, and exploring the robustness of the recommended procedure with different replicate variance estimation procedures.

ACKNOWLEDGEMENTS

The authors would like to thank Elizabeth Huang and James Fagan of the U.S. Census Bureau, two anonymous referees, and the associate editor for their helpful comments on earlier versions of this manuscript, and J.N.K Rao for his useful comments on the original simulation study. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

DeGROOT, M. (1986). *Probability and Statistics*. Reading, MA: Addison-Wesley Publishing, Inc.

FAY, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

FAY, R.E. (1995). VPLX: Variance Estimation for Complex Surveys. Program Documentation: Unpublished Bureau of the Census Report.

HOAGLIN, D.C., and IGLEWICZ, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 83, 1147-1149.

JUDKINS, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, 25-45.

KOVACEVIC, M., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization – An empirical Study. *Survey Methodology*, 23, 41-52.

LUERY, D.M. (1990). Survey of Construction Technical Paper. Unpublished draft Bureau of the Census internal documentation.

NAYLOR, T.H., BALINTFY, J.L., BURDICK, D. S., and CHU, K. (1968). *Computer Simulation Techniques*. New York: John Wiley and Sons, Inc.

RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

RAO, J.N.K., and SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

THOMPSON, K.J. (1998). Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC). Technical Report #ESM-9801, available from the Economic Statistical Methods and Programming Division.

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

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 163-172
Statistics Canada

The Impact of Different Rotation Patterns on the Sampling Variance of Seasonally Adjusted and Trend Estimates

C.H. McLAREN and D.G. STEEL[1]

ABSTRACT

Many economic and social time series are based on sample surveys which have complex sample designs. The sample design affects the properties of the time series. In particular, the overlap of the sample from period to period affects the variability of the time series of survey estimates, and the seasonally adjusted and trend estimates produced from them. The Census X11 and X11ARIMA packages are commonly used to produce seasonally adjusted estimates and can also be used to produce estimates of trend. This paper considers the implications of different overlap patterns on the sampling variance of seasonally adjusted and trend estimates obtained from time series based on sample surveys.

KEY WORDS: X11; X11ARIMA; Seasonal adjustment; Trend estimation; Rotation patterns.

1. INTRODUCTION

Many important time series are based on repeated sample surveys which have complex patterns of sample overlap from period to period. The use of sampling means that the estimated time series have a component of variability due to sampling errors and for many series this will be a major source of variability. The sample design, in particular the overlap pattern, affects the variability of the time series of survey estimates.

Increasingly, analysis of time series is concentrating on assessing underlying patterns of change or trends based on analysis of the seasonally adjusted series. Most government statistical agencies have calculated seasonally adjusted series for many years. Kenny and Durbin (1982) noted that policy analysts frequently say that they are more interested in underlying trends than following irregular fluctuations in the de-seasonalized monthly values. A similar view is expressed by Smith (1997). For more than 10 years the Australian Bureau of Statistics (ABS) has published series of trend estimates obtained by applying Henderson Moving Averages (HMAs) (Henderson 1916) to the seasonally adjusted series to smooth out the irregular components of the series (ABS 1987). Other government statistical agencies also produce trend estimates using a variety of methods (Knowles 1997). Since seasonally adjusted and trend estimates are obtained by processes applied to the original series, they are also influenced by sampling errors. Bell and Kramer (1999) note that the variance of seasonally adjusted estimates will often be dominated by the contribution from sampling error. Some series are based on independent samples over time, but usually the samples used have a degree of overlap from period to period to reduce costs and the standard errors of estimates of change between two consecutive time periods (Kish 1998).

A key issue in the development of the design of a repeated survey is the rotation pattern, that is, the pattern of a selected unit's inclusion in the survey over time, which will determine the sample overlap. The aim of this paper is to determine the effects of the rotation pattern used on the sampling variance of the estimated seasonally adjusted and trend series obtained using the Census X11 method developed by Shiskin, Young and Musgrave (1967) and X11ARIMA developed by Dagum (1980 and 1988). We will focus on the estimates of the level and one period change in the seasonally adjusted and trend estimates.

2. ROTATION PATTERNS

Consider a univariate time series with values y_t, $t = 1,...T$, obtained from a repeated sample survey. The observed value at time t is related to the true value of the series in the finite population, Y_t, by

$$y_t = Y_t + e_t$$

where e_t is the sampling error. The series Y_t is thought to consist of trend-cycle, seasonal and irregular components T_t, S_t and I_t, so that

$$y_t = T_t + S_t + I_t + e_t.$$

In some cases a multiplicative decomposition may be more appropriate. Many statistical agencies produce seasonally adjusted series by attempting to estimate S_t and remove it from the series, usually using some combination of linear filters. Most commonly used is the Census X11 method developed by Shiskin *et al.* (1967) and X11ARIMA developed by Dagum (1980 and 1988). Findley, Monsell, Otto, Bell and Pugh (1998) described further enhancements embodied in X12ARIMA. The ABS also publishes trend

[1] C.H. McLaren and D.G. Steel, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia. E-mail: craigmc@uow.edu.au, dsteel@uow.edu.au.

estimates obtained by applying HMAs to the seasonally adjusted series and encourages users to base their interpretation of the series on these trend estimates (Linacre and Zarb 1991; ABS 1993). The HMAs were originally derived by Henderson (1916) for use in actuarial work and are used within X11, X11ARIMA and X12ARIMA to de-trend series for seasonal adjustment purposes. Kenny and Durbin (1982) and Gray and Thomson (1996) explain the derivation of the HMAs. Users can also produce trend estimates by applying filters to the published seasonally adjusted estimates. Kenny and Durbin (1982) noted that there is no unique definition of trend and that different filters may be used according to the degree of smoothness and sensitivity required. Knowles and Kenny (1997) investigated methods of trend estimation for official statistical series. For monthly series they recommended the use of HMAs, with the length of the filter being 13 or 23 depending on the volatility of the series in question.

The autocorrelation structure of the observed series is determined by the autocorrelation of the series Y_t and e_t, which will then affect the estimates of the trend, seasonally and irregular components. The covariance structure of the sampling error series, e_t, can be estimated from the unit level survey data. By obtaining such estimates, it is possible to obtain estimates of the sampling variance of the estimated trend, seasonally adjusted and irregular series. Various methods for doing this have been proposed; for example Steel and DeMel (1988) considered the effect of linear filters on the spectrum of the sampling error series and Wolter and Monsour (1981) used an approach based on the effect of linear filters on the autocovariance function. Sutcliffe (1993) adopted a similar approach using a linear approximation to the X11 procedure. Pfeffermann (1994) proposed a method which develops an estimate of sampling error directly from the estimated time series using various simplifying assumptions. These approaches do not explicitly model the time series. Other authors, for example Bell and Wilcox (1993), Tiller (1992), Burridge and Wallis (1985) and Hausman and Watson (1985), considered explicit ARIMA models for both the true series and the sampling error series, and concentrated on the estimation of the parameters of the models. These papers do not consider the effect of different rotation patterns and concentrate on producing estimates of the variances of seasonally adjusted estimates for the particular rotation pattern used.

The rotation pattern used in the survey will affect the autocorrelation structure of the sampling error series and hence the sampling variance of the original, seasonally adjusted and trend estimates. Several considerations are taken into account in deciding upon a rotation pattern. High sample overlap between consecutive periods reduces the sampling variance of estimates of change between the periods and high sample overlap between periods 12 months apart reduces the sampling variance of estimates of annual change. The first occasion that a selected unit is included in the survey is usually the most expensive. By keeping selected units in the survey for longer the cost of the survey is reduced. This leads to rotation patterns in which a selected unit is included every period for as long as possible. However, a selected unit must eventually be rotated out of the survey. Besides the ethical consideration of spreading respondent load, there is the possible deterioration in response rate and quality of data reported if the same unit is included for a large number of occasions (see Kalton and Citro 1993, for a discussion of these issues).

Rotation patterns vary in terms of the number of times a unit is included in the survey and the time interval between inclusions. We concentrate on monthly labour force surveys (MLFSs). The rotation patterns used in practice are special cases of the *a-b-a(m)* rotation patterns where selected units are included for a consecutive months, removed from the survey for b months then re-included for a further a months. The pattern is repeated so that selected units are included for a total of m occasions. Rao and Graham (1964) considered the estimation of the finite population means and totals for this class of rotation patterns. The United States Current Population Survey (CPS) uses a 4-8-4(8) pattern (Fuller, Adam and Yansaneh 1992). Putting $b = 0$ gives an *in-for-m* rotation pattern in which selected dwellings are included for m months after which they are removed from the sample. The case $m = 6$ corresponds to the Canadian rotation pattern (Singh, Drew, Gambino and Mayda 1990) and $m = 8$ corresponds to the Australian pattern (ABS 1992). Steel (1997) noted that the British quarterly labour force survey approximately corresponds to a monthly survey with a 1-2-1(5) rotation pattern.

We consider the sampling variance of the seasonally adjusted and trend estimates associated with the rotation patterns currently used in MLFSs and a number of rotation patterns that, while not currently used, may have some desirable properties. This will give an indication of which rotation patterns are better in terms of the component of the variability of the estimated series that is affected by the sample design.

3. SAMPLING VARIANCE OF SEASONALLY ADJUSTED AND TREND ESTIMATES

Let y_T be the vector containing the values of the time series of survey estimates up to time T and Y_T be the vector containing the true population values. The sampling variance of the original series is denoted by $V(y_T | Y_T)$. Consider a linear filter which is used to obtain values from y_T by applying a vector of filter weights w_t. The filter weights are non-random and have no connection with the survey weights used in calculating the survey estimates y_t. The filter weight vectors w_t depends on the time period for which the filtered value refers. The weights are constant within the body of the series but may be modified at the beginning and end. The filtered value at time t is

$$\tilde{y}_t = w_t' y_T. \qquad (1)$$

Then

$$V(\bar{y}_t \mid \boldsymbol{Y}_T) = \boldsymbol{w}_t' V(\boldsymbol{y}_T \mid \boldsymbol{Y}_T)\boldsymbol{w}_t \qquad (2)$$

is the sampling variance of the filtered value at time t. The sampling error of the filtered value is the difference between $\boldsymbol{w}_t'\boldsymbol{y}_T$ and $\boldsymbol{w}_t'\boldsymbol{Y}_T$, which is conditional on the values of the true series, $\boldsymbol{Y}_T$. This is the difference between the filtered value obtained from the series of estimates ending at time T and the value that would be obtained if that series was observed without sampling error. We focus on this component as it is the sampling variance that can be altered by changing the sample design. The variance associated with $\boldsymbol{Y}_T$ has not been taken into account. Wolter and Monsour (1981) discussed the issue of total variance versus sampling error variance. There may be advantages in considering the total variance in interpreting the resulting series but when we are considering sample design issues, such as the choice of rotation pattern, we focus on the component that is directly affected by decisions made about the sample design. If the sampling error does not contribute significantly to the variability of the series then decisions about the sample design are not as important as they are when the sampling error is a major contributor, although it still seems sensible to use as effective a sample design as possible.

To determine the effect of different rotation patterns on the sampling variance of a particular filtered series, we need an estimate of $V(\boldsymbol{y}_T \mid \boldsymbol{Y}_T)$ for different rotation patterns. Previous work on estimating variances of seasonally adjusted series has either ignored the rotation pattern and assumed independent samples at each time point, or taken it as fixed and used an estimation method that takes it into account. We need a model for $V(\boldsymbol{y}_T \mid \boldsymbol{Y}_T)$ that reflects the effect of the different rotation patterns that could be used.

The analysis of the effect of different rotation patterns is simplified if the series of sampling errors has a stable autocorrelation structure. The precise form of the autocorrelation function will depend on the series and should reflect the complexities of the design. For example Steel and DeMel (1988) suggested a model for the Australian Monthly Labour Force data and Bell and Wilcox (1993) suggested a model for the United States Retail Trade series. Bell and Hillmer (1990) and Miazaki and Dorea (1993) also considered modelling of survey errors by time series models. Dempster and Hwang (1993) and Lee (1990) considered approaches to estimating and modelling sampling error correlations for the US CPS.

Our approach is to assume that the series of sampling errors, e_t has constant variance. A model is needed for the correlation between the sampling errors of y_t and y_{t+s}. All the rotation patterns considered imply that the sample at any particular time will consist of a number of panels. A panel is a set of units that are included and removed from the survey at the same time. When a panel is rotated out of the survey it will be replaced by another panel. The set of panels related in this way is referred to as a rotation group. Most MLFSs use multistage sampling and when a panel is rotated out of the survey it is replaced by another panel of nearby households (see ABS 1992; Singh *et al.* 1990). Hence it is assumed that the sampling correlation between estimates obtained from the same rotation group s periods apart is $r(s)$ if no rotation has occurred and $d(s)$ if rotation has occurred. We will assume that the estimate at time t is, at least approximately, the average of estimates from each rotation group and that estimates from different rotation groups, which will usually be in different PSUs and spatially well separated, are independent.

These assumptions imply that the sampling correlation between y_t and y_{t+s} is

$$R(s) = d(s) + k(s)(r(s) - d(s)) \qquad (3)$$

where $k(s)$ is proportion of the sample in common between the two time periods. The sample overlap factor $k(s)$ is determined by the rotation pattern. For example, for an *in-for-m* rotation pattern $k(s) = 1 - s/m, s = 0, ..., m - 1$ and zero otherwise, assuming that the same number of dwellings are added and dropped from the sample each month. If different panels in the same rotation group are independent, then $d(s) = 0$, but in general this will not be the case. This model is essentially the same as derived by Scott, Smith and Jones (1977). An example of an *in-for-4* rotation pattern over an eight month period is illustrated in Table 1. Different panels are denoted by different letters and the subscript indicates the number of times the panel has been included in the survey up to the time period indicated.

Table 1
Structure of *in-for-4* Rotation Pattern

	Time Period							
Rotation Group	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$
1	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
2	c_4	d_1	d_2	d_3	d_4	e_1	e_2	e_3
3	f_3	f_4	g_1	g_2	g_3	g_4	h_1	h_2
4	i_2	i_3	i_4	j_1	j_2	j_3	j_4	k_1

In this case $r(2)$ is the correlation arising from say a_2 and a_4, whereas $d(2)$ is the correlation associated with a_3 and b_1. Binder and Hidiroglou (1988) and Fuller *et al.* (1992) provided discussions of the data structure implied by some other rotation patterns.

The assumption that the variance of the sampling error series is constant implies that no major changes to the sample design or the population structure occur, at least over the effective length of the filters being considered. The assumption of stable autocorrelations, $r(s)$ and $d(s)$, for the population correlation also implies no major changes to the sample design or population. Estimates for $r(s)$ and $d(s)$ in (3) were obtained from a study by Bell (1998). The values used are from the Australian Labour Force Survey (ALFS) for the proportion of persons employed and also the proportion of persons unemployed and are shown in

Table 2. These were obtained by treating the rotation groups in the ALFS as replicates and measuring the autocorrelation at the rotation group level. A model given in Bell (1998) was used to extrapolate values beyond the given lags.

Table 2
Autocorelations – ALFS

Proportion of employed persons								
lag	1	2	3	4	5	6	7	8
$r(s)$	0.80	0.71	0.64	0.57	0.50	0.45	0.40	0.36
$d(s)$	0.15	0.15	0.14	0.13	0.12	0.11	0.11	0.10
Proportion of unemployed persons								
	1	2	3	4	5	6	7	8
$r(s)$	0.62	0.52	0.44	0.37	0.31	0.26	0.22	0.19
$d(s)$	0.11	0.11	0.10	0.09	0.09	0.08	0.08	0.07

Sutcliffe and Lee (1995) studied the standard errors of seasonally adjusted and trend estimates of level and movement under a small number of different rotation patterns. They assumed a simple geometric decay model for the correlations between survey estimates with a population correlation of $\rho = 0.8$, *i.e.*, $R(s) = \rho^s$, which decreases more rapidly than the values given in Table 2.

4. LINEAR APPROXIMATIONS FOR SEASONALLY ADJUSTED AND TREND ESTIMATES

The X11 method consists of an iterative application of moving averages resulting in a symmetric filter for the central values, and asymmetric filters for the values at the beginning and end of the series. The final seasonally adjusted and trend estimates produced by X11 can be approximated by linear filters. Several authors; for example, Young (1968), Cleveland and Tiao (1976), Wallis (1982), and Sutcliffe (1993), have produced linear approximations to the X11 procedure. The X11ARIMA procedure (Dagum 1980, 1988) is an extension of X11 and extrapolates the original series at both ends by an ARIMA model. The effect of the ARIMA extrapolation can be incorporated into the filter weights and these weights can be applied to the data alone. Dagum, Chhab and Chiu (1996) considered a Cascade method approach, where the Cascade filters are a result of the convolution of the various predetermined linear filters used within both X11 and X11ARIMA. We used this approach to realistically approximate both the X11 and X11ARIMA procedures.

Define the matrix whose rows contain the filter weights of 13 term HMAs for both symmetric and asymmetric filters as $\mathbf{H}_{13}$. The matrix of weights corresponding to a 3×3 moving average (ma) is denoted as $\mathbf{S}_{3\times3}$ and that corresponding to a 3×5 ma is denoted as $\mathbf{S}_{3\times5}$. These are used for estimation of seasonal factors. The matrix $\mathbf{D}$ is defined as a 12 term centered ma and $\mathbf{I}$ is an identity matrix. The notation c indicates the complement of a filter, for example $\mathbf{D}^c = \mathbf{I} - \mathbf{D}$. The Seasonal Adjustment Cascade filters are written as

$$\mathbf{S} = \mathbf{I} - \mathbf{D}^c\mathbf{S}_{3\times5}[\mathbf{H}_{13}(\mathbf{D}^c\mathbf{S}_{3\times3}\mathbf{D}^c)^c]^c.$$

The trend Cascade filters used for the estimation of trend are then found by multiplying the seasonally adjusted filter by a trend filter. At the end of the series the Cascade filters for trend and seasonally adjusted estimates will differ according to whether X11 or X11ARIMA is used.

We consider the following different combinations of the internal filters of X11 and X11ARIMA:

1. Standard X11 Cascade filter: This corresponds to a 13 term HMA for estimation of trend (H13), 3×3 ma for the first estimation of the seasonal factors ($S1_{3\times3}$), 3×5 ma for estimation of seasonal factors ($S2_{3\times5}$), and no modification for outliers.
2. Standard X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H13, $S1_{3\times3}$, $S2_{3\times5}$, and extended forecasts from an ARIMA model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.4B)(1 - 0.6B^{12})a_t$, where B is the backward shift operator and a_t is a white noise process, and no modification for outliers.
3. Short X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H9, $S1_{3\times3}$, $S2_{3\times5}$, and extended forecasts from a model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.3B)(1 - 0.3B^{12})a_t$, and no modification for outliers.
4. Long X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H23, $S1_{3\times3}$, $S2_{3\times5}$, and extended forecasts from a model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.8B)(1 - 0.8B^{12})a_t$, and no modification for outliers.

Combinations 2 and 3 have been observed by Dagum (1983) to be applicable in a number of cases. The linear approximations chosen allow us to examine the effect of different rotation patterns for a range of filters used in practice, which involve HMAs of different lengths.

For each combination of filters the corresponding Cascade filter provides a vector of filter weights for the seasonally adjusted estimates and a different weight vector for the final trend estimates. These can then be substituted into equation (2) to obtain the sampling variances for a particular rotation pattern by using the appropriate values for $V(y_T|Y_T)$. When computing change estimates the data vector y_T remains unchanged and the weights that are applied change. For example, $w_{t+1} - w_t$ can be used for a one month difference. This basic approach is the same as that adopted by Wolter and Monsour (1981) who proposed estimating the variance of seasonally adjusted estimates using (2) with weights chosen that reasonably approximate the seasonal adjustment process and using a survey based estimate of $V(y_T|Y_T)$. We also consider trend filters and different realisations of X11ARIMA and rotation patterns.

The X11ARIMA models considered in this paper are representative of those commonly used in practice.

Additional complications arise from the use of ARIMA forecasts in the X11ARIMA approach. For example, we assume no misspecification of the ARIMA model. The ARIMA model is typically identified and estimated using previous survey data. The sampling error for previous time points could influence the choice of ARIMA model and X11 filters. This could be taken into consideration by modification of the variance in (2).

The initial trend and seasonally adjusted estimates for time t will be made using the time series of estimates ending at time t, that is y_t, giving the filtered value $w_t'y_t$. The value that would be obtained if there was no sampling error is $w_t'Y_t$. The sampling error considered in this paper is $w_t'y_t - w_t'Y_t$. As estimates are added to the series the filtered value for time t may change, but there will come a time point, $t+s$, after which there is no appreciable change. The final filtered value for time t based on the survey estimates can be written as $w_t^{*\prime}y_{t+s}$, for a final symmetric weight vector w_t^*. Similarly the final value that would be obtained if there were no sampling error would be $w_t^{*\prime}Y_{t+s}$. Bell and Kramer (1999) considered the difference $w_t'y_t - w_t^{*\prime}Y_{t+s}$, which includes the forecast error. This difference can be decomposed as

$$w_t'y_t - w_t^{*\prime}Y_{t+s} = (w_t'y_t - w_t'Y_t) + (w_t'Y_t - w_t^{*\prime}Y_{t+s}).$$

We have considered how different rotation patterns affect the first term in this decomposition. The second term involves the series observed without sampling error and is unaffected by the sample design, including the rotation pattern. Bell and Kramer (1999) considered the series of US Housing Starts involving five or more units and showed that the total variance of the trend series showed large increases at the end of the series due to forecasting errors. This is due to the revisions in the initial trend estimates that are made as estimates are added to the series. Steel and McLaren (2000) considered the effect of different rotation patterns on the observed revision of the initial trend estimates, which is $w_t'y_t - w_t^{*\prime}y_{t+s}$. They noted that the relative importance of the component due to sampling error will depend on how the true series is evolving around the period being considered.

5. RESULTS

We use filters corresponding to the level and one month difference for both the seasonally adjusted and trend estimates at the very end of the series. Tables 3 to 6 summarise the effect of different rotation patterns for each Cascade filter combination. These tables give, for a selection of rotation patterns, the ratio of the sampling variance of the estimates under consideration divided by the sampling variance that would be obtained when there is complete rotation each month. The ratios obtained in the middle of the series give the same general conclusions (McLaren 1999).

Table 3
Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 1)

Rotation Pattern	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	0.99	0.99	0.99	1.00	0.99	1.00	0.68	0.79
1-2-1(8)	0.98	0.99	0.97	0.99	0.98	0.99	0.64	0.77
1-1-1(6)	1.01	1.01	1.00	1.00	1.17	1.14	0.7	0.82
2-2-2(8)	1.02	1.02	0.61	0.71	1.26	1.23	0.83	0.95
2-10-2(4)	1.04	1.04	0.61	0.71	1.35	1.30	1.32	1.26
3-3-3(6)	1.07	1.06	0.48	0.61	1.52	1.44	1.29	1.25
4-8-4(8)	1.10	1.08	0.42	0.57	1.69	1.57	1.42	1.34
6-6-6(12)	1.10	1.08	0.36	0.52	1.76	1.64	1.22	1.22
in-for-6	1.10	1.08	0.36	0.52	1.76	1.64	1.22	1.22
in-for-8	1.09	1.08	0.33	0.50	1.78	1.65	1.06	1.13
no rotation	1.08	1.08	0.24	0.44	1.80	1.69	0.75	0.95

Table 4
Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 2)

Rotation Pattern	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	1.01	1.01	0.99	1.00	1.06	1.05	0.69	0.80
1-2-1(8)	1.00	1.00	0.96	0.99	1.07	1.05	0.66	0.78
1-1-1(6)	1.04	1.03	1.00	1.00	1.22	1.17	0.65	0.77
2-2-2(8)	1.05	1.04	0.60	0.71	1.32	1.26	0.81	0.92
2-10-2(4)	1.02	1.03	0.60	0.71	1.26	1.23	1.19	1.17
3-3-3(6)	1.08	1.06	0.49	0.61	1.49	1.40	1.19	1.16
4-8-4(8)	1.06	1.06	0.41	0.56	1.56	1.47	1.13	1.13
6-6-6(12)	1.08	1.07	0.35	0.52	1.67	1.56	0.93	1.01
in-for-6	1.10	1.08	0.36	0.52	1.69	1.56	0.94	1.01
in-for-8	1.11	1.08	0.32	0.49	1.75	1.61	0.82	0.93
no rotation	1.14	1.11	0.24	0.43	1.89	1.73	0.59	0.78

Table 5
Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 3)

Rotation Pattern	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	0.99	0.99	0.96	0.98	0.99	0.99	0.68	0.79
1-2-1(8)	0.97	0.99	0.93	0.97	0.98	0.99	0.64	0.77
1-1-1(6)	1.04	1.02	0.99	0.99	1.11	1.08	0.6	0.72
2-2-2(8)	1.07	1.06	0.60	0.71	1.23	1.19	0.89	0.95
2-10-2(4)	1.05	1.06	0.61	0.72	1.21	1.20	1.07	1.08
3-3-3(6)	1.15	1.12	0.51	0.63	1.41	1.32	1.02	1.02
4-8-4(8)	1.12	1.11	0.44	0.58	1.41	1.35	0.85	0.93
6-6-6(12)	1.14	1.13	0.37	0.53	1.47	1.39	0.69	0.82
in-for-6	1.16	1.13	0.38	0.53	1.49	1.40	0.70	0.81
in-for-8	1.17	1.14	0.34	0.51	1.52	1.42	0.61	0.76
no rotation	1.22	1.17	0.25	0.44	1.62	1.50	0.44	0.64

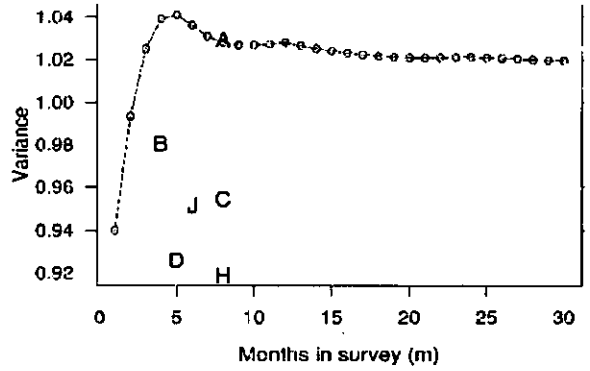
Table 6
Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 4)

Rotation Pattern	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	1.02	1.02	0.99	1.00	1.25	1.19	0.75	0.87
1-2-1(8)	1.02	1.02	0.97	0.99	1.28	1.21	0.7	0.84
1-1-1(6)	1.06	1.04	1.00	1.00	1.49	1.39	0.92	1.01
2-2-2(8)	1.06	1.04	0.60	0.71	1.57	1.47	0.98	1.09
2-10-2(4)	1.00	1.01	0.60	0.70	1.30	1.27	1.49	1.37
3-3-3(6)	1.07	1.05	0.48	0.61	1.64	1.54	1.34	1.36
4-8-4(8)	1.05	1.04	0.41	0.56	1.73	1.63	1.92	1.69
6-6-6(12)	1.08	1.06	0.35	0.51	2.00	1.84	1.87	1.68
in-for-6	1.09	1.07	0.35	0.52	2.00	1.84	1.90	1.70
in-for-8	1.11	1.08	0.32	0.49	2.15	1.96	1.73	1.62
no rotation	1.17	1.12	0.24	0.43	2.56	2.27	1.11	1.33

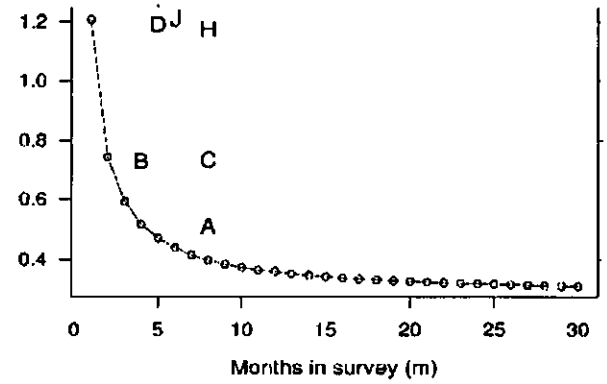
5.1 X11 – Concurrent Standard Cascade Filters

The results using the standard X11 filters (combination 1) are shown in Table 3. Figures 1(a) to 1(d) show the sampling variance of the level and one month difference for the seasonally adjusted and trend estimates at the end of the series divided by the variance of the original estimate of level plotted against the total number of times a selected unit is included. Results for the variable employment have been plotted for selected *a-b-a(m)* patterns and the *in-for-m* rotation patterns for *m* going from 1 to 30. An *in-for*-30 rotation pattern is indicative of having no rotation.

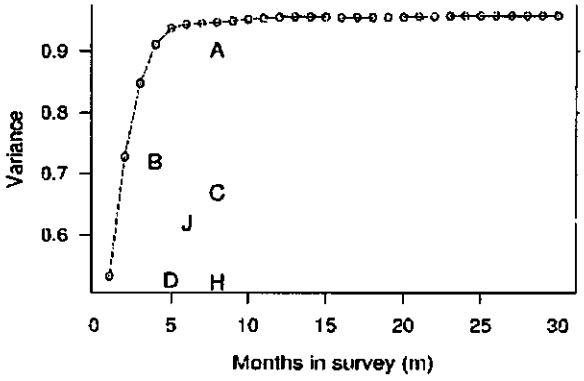
Columns 1 and 2 in Table 3 show that for the variance of the seasonally adjusted level estimates, rotation patterns with no monthly overlap perform well. Using rotation patterns with annual overlap did not help appreciably. However, for the one month change in seasonally adjusted estimates, the benefit of having high monthly overlap becomes evident (see Figure 1(b) and columns 3 and 4 of Table 3). The variances associated with the *in-for-m* rotation patterns are effectively a function of $1/m$, the proportion of the sample that does not overlap. Those rotation patterns used in Canada and Australia perform well. The best option is no rotation but, as discussed in section 2, this is not a practical option. Figures 1(a) and 1(b) show that rotation patterns that have the same degree of monthly sample overlap have similar variances for estimates of the level and one month change in the seasonally adjusted series.



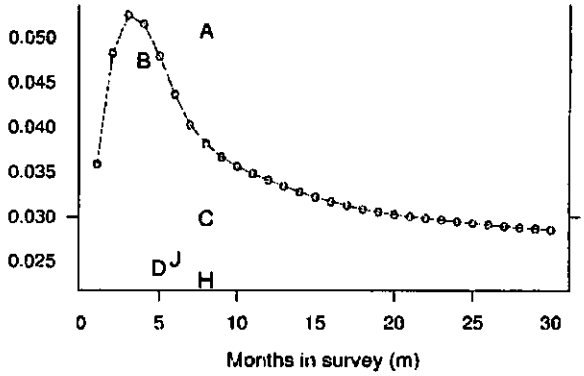
1a) Seas. adjusted: level estimates



1b) Seas. adjusted: one month change



1c) Trend: level estimates



1d) Trend: one month change

Figure 1. Ratio of the sampling variance to the variance of the original series for chosen rotation patterns for combination 1 (X11) for the variable employment where A = 4-8-4(8), B = 2-10-2(4), C = 2-2-2(8), D = 1-2-1(5), H = 1-2-1(8), J = 1-1-1(6).

For the level of trend estimates the variance increases as the amount of monthly sample overlap increases (see Figure 1(c) and columns 5 and 6 of Table 3). For the *in-for-m* rotation patterns there is a rapid increase in variance as m goes from 1 to 5. The rotation patterns of 1-2-1(5) and 1-2-1(8) perform as well as having an independent sample each month and considerably better than rotation patterns that involve monthly overlap. This is primarily due to the fact that for a moving average, it is better to average over independent observations than positively correlated ones. The larger variance of the 1-1-1(6) pattern compared with that of 1-2-1(5) and 1-2-1(8) suggest that, for those patterns with no monthly overlap, the interval between the re-inclusion of units in the sample has some effect.

Figure 1(d) and columns 7 and 8 of Table 3, show that for one month changes in trend estimates the variance increases very rapidly as m increases from 1 to 3 and decreases rapidly as m increases from 4. The *in-for*-3 rotation pattern seems to be the worst among those considered, and the currently used rotation patterns can be significantly improved upon. For example, using a 1-2-1(8) instead of a 4-8-4(8) rotation pattern would reduce the variance in the one month change in trend estimates for employment by 55 percent and 43 percent for unemployment. While the degree of monthly overlap is still a key factor, the pattern of inclusion also plays a role, for example the 2-2-2(8) pattern has lower variance than the *in-for*-2 or 2-10-2(4) patterns. Moreover, for one month changes in the trend estimates the best performing rotation patterns are 1-2-1(5) and 1-2-1(8) which perform considerably better than using complete rotation each month. This result arises because one month changes in trend estimates effectively look at differences in the seasonally adjusted series a few months apart and the 1-2-1(m) rotation patterns lead to positive correlations between estimates 3 months apart. Similar results were obtained in a study by McLaren and Steel (1997) using Sutcliffe's (1993) approximation to X11.

The results show that for the estimation of the current level of trend and the latest movement in trend, the 1-2-1(m) rotation patterns give considerably lower sampling variances than the rotation patterns currently in use.

5.2 X11ARIMA – Concurrent Cascade Filters with Extrapolations

Results for the filter combinations 2, 3 and 4 are given in Tables 4, 5 and 6 respectively. Figures 2(a) to 2(d) present results for combination 4 for employment.

Columns 1 and 2 of Tables 4, 5 and 6 show that rotation patterns with low monthly overlap perform almost as well as complete rotation for seasonally adjusted level estimates. Rotation patterns with high monthly overlap have higher variances, particularly for combination 3 which corresponds to the use of the 9 term HMA.

There is minimal difference between the ratios of the four different combinations for the one month change in the seasonally adjusted estimates (columns 3 and 4 in all tables). Rotation patterns with high monthly sample overlap still perform better than those with low or no monthly overlap regardless of the X11/X11ARIMA combination used.

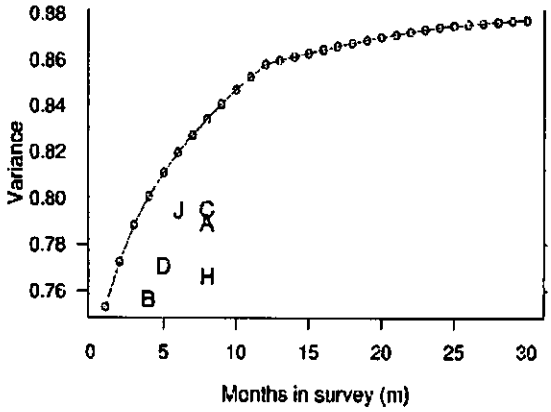
For the level of trend estimates, rotation patterns with a higher degree of sample overlap again have a greater variance ratio. The 1-2-1(5) and 1-2-1(8) rotation patterns still out-perform the other rotation patterns for each combination of filters, although they do not perform as well as an independent sample for combinations 2 and 4.

For one month changes in the trend estimates the better performing rotation patterns are again 1-2-1(5) and 1-2-1(8) which perform better than the independent sample for all four combinations of filters. For combination 3, rotation patterns with high monthly overlap perform equally as well as the 1-2-1(m) rotation patterns. For combinations 2 and 3 the 1-1-1(6) pattern is slightly better than the 1-2-1(m) patterns. Substantial improvements over the currently used rotation patterns can be achieved by using 1-2-1(m) rotation patterns. For example, for the employment variable, changing from an 4-8-4(8) to a 1-2-1(8) would produce gains of 42, 25 and 64 percent using combinations 2, 3 and 4, respectively.

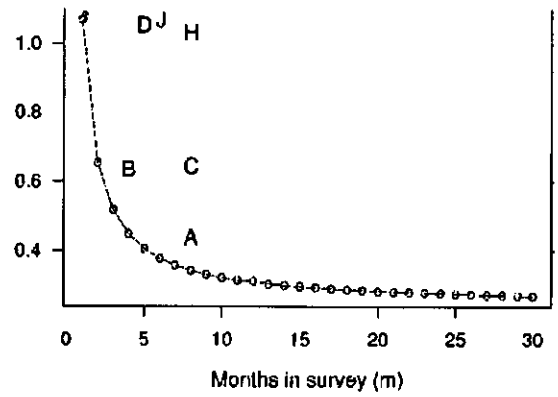
These results are based on the ALFS correlation estimates which, being based on survey estimates, will be subject to sampling error. The trend filters considered are not derived using these estimates. The same general conclusions concerning the impact of different rotation patterns are obtained for the two correlation models which use reasonably different correlations. We believe that the conclusions will apply for the range of correlation models contained between these two models. Similar conclusions are also obtained by McLaren and Steel (1997) using a correlation model derived by Steel (1996) for UK employment and unemployment.

6. DISCUSSION

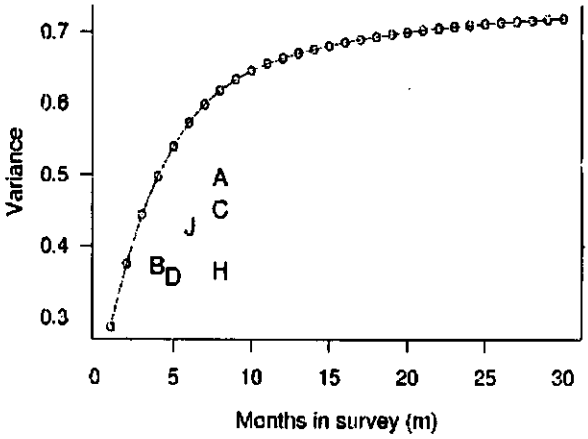
The rotation patterns currently used, such as *in-for*-8, *in-for*-6 and 4-8-4(8), are sensible if the one month change in seasonally adjusted estimates are the key statistics to be analysed. We believe that examination of the one month change in seasonally adjusted estimates is often not a reliable way of assessing current trends. It is necessary to look at the pattern of change over recent months. This can be done using filters to obtain an estimate of the trend. The results here suggest if the main use of the survey is to provide an assessment of trend then quite different rotation patterns should be used. Specifically, the 1-2-1(m) rotation patterns performed well for reducing the variance of the level of trend estimates and the difference between two consecutive trend estimates for a range of different filter combinations. The 1-2-1(m) rotation patterns also performed well for the sampling variance of the seasonally adjusted level estimates. Hence, in designing the rotation pattern for a repeated survey, the relative importance of



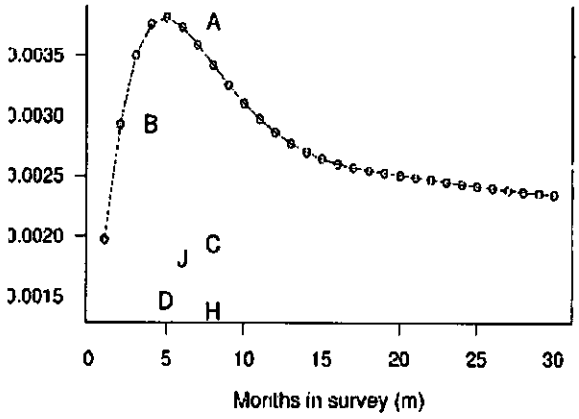
2a) Seas. adjusted: level estimates



2b) Seas. adjusted: one month change



2c) Trend: level estimates



2d) Trend: one month change

Figure 2. Ratio of the sampling variance to the variance of the original series for chosen rotation patterns for Combination 4 (X11ARIMA) for the variable employment where A = 4-8-4, B=2-10-2(4), C=2-2-2(8), D=1,2-1(5), H=1-2-1(8), J=1-1-1(6).

seasonally adjusted and trend estimates needs to be carefully considered. Examining Figures 1 and 2 shows that the rotation pattern 2-2-2(8), is a reasonable compromise if the level and one months change in seasonally adjusted and trend estimates are both considered important. Bell (1999) also considered the effect of four different rotation patterns on the sampling variance of the level and one month change in the original, unadjusted, estimates and also trend estimates obtained using X11 and a 13 point HMA. He also identifies the 2-2-2(8) rotation pattern as a compromise design.

Even if analysts do not formally use trend estimates, the assessment of trend will involve looking at changes in seasonally adjusted estimates a few months apart. McLaren (1999) gives results which show that the 1-2-1(m) rotation patterns will be suitable if the assessment of trends involve looking at changes in seasonally adjusted estimates over 3 or 6 months. The results also suggest that such rotation patterns perform well for estimates of the change in trend estimates over the most recent 3 and 6 months.

The evaluation criterion used in this paper is the sampling variance of the trend and seasonally adjusted estimates, which is the factor affected by the sample design. Steel and McLaren (2000) considered assessing different rotation patterns in terms of the degree of revisions of these estimates at the end points and reached similar conclusions regarding the rotation patterns.

ACKNOWLEDGEMENT

This research was supported by the Australian Research Council and the Australian Bureau of Statistics (ABS). The views expressed in this paper may not necessarily reflect the views of either organisation. We would like to thank the associate editor and the referees for their comments and Geoff Lee, Andrew Sutcliffe and Phillip Bell from the ABS, and Norma Chhab from Statistics Canada.

REFERENCES

AUSTRALIAN BUREAU OF STATISTICS (1987). *A Guide to Smoothing Time Series – Estimates of "Trend"*. Australian Bureau of Statistics, catalogue no. 1316.0, Canberra.

AUSTRALIAN BUREAU OF STATISTICS (1992). *Information Paper: Labour Force Survey Sample Design.* Australian Bureau of Statistics, catalogue no. 6269.0, Canberra.

AUSTRALIAN BUREAU OF STATISTICS (1993). *A Guide to Interpreting Time Series – Monitoring "Trends", An Overview.* Australian Bureau of Statistics, catalogue no. 1348.0, Canberra.

BELL, P.A. (1998). *Using State Space Models and Composite Estimation to Measure the Effects of Telephone Interviewing on Labour Force Estimates.* Working Papers in Econometrics and Applied Statistics, No. 98/2, Australian Bureau of Statistics, catalogue no. 1351.0, Canberra.

BELL, P.A. (1999). *The Impact of Sample Rotation Patterns and Composite Estimation on Survey Outcomes.* Working Papers in Econometrics and Applied Statistics, No. 99/1, Australian Bureau of Statistics, catalogue no. 1352.0, Canberra.

BELL, W.R., and HILLMER, S. (1990). Time series methods for survey estimation. *Survey Methodology*, 16, 195-215.

BELL, W.R., and KRAMER, M. (1999). Towards variances for X-11 seasonal adjustment. *Survey Methodology*, 25, 13-29.

BELL, W.R., and WILCOX, D.W. (1993). The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics*, 555, 235-265.

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

BURRIDGE, P., and WALLIS, K.F. (1985). Calculation of seasonally adjusted series. *Journal of the American Statistical Association*, 80, 541-552.

CLEVELAND, W.P., and TIAO, G.C. (1976). Decomposition of seasonal time series: a model for the X-11 program. *Journal of the American Statistical Association*, 71, 581-587.

DAGUM, E.B. (1980). *The X-11ARIMA Seasonal Adjustment Method.* Catalogue no. 12-564E, Statistics Canada, Ottawa.

DAGUM, E.B. (1983). Spectral properties of the concurrent and forecasting seasonal linear filters of the X-11ARIMA method. *The Canadian Journal of Statistics*, 11, 73-90.

DAGUM, E.B. (1988). *The X-11ARIMA/88 Seasonal Adjustment Methods – Foundations and User's Manual.* Statistics Canada, Ottawa.

DAGUM, E.B., CHHAB, N., and CHIU, K. (1996). Derivation and properties of the X-11ARIMA and Census X-11 linear filters. *Journal of Official Statistics*, 12, 329-347.

DEMPSTER, P.A., and HWANG, J.-S. (1993). Component models and Bayesian technology for estimation of state employment and unemployment rates. *Proceedings of the Bureau of the Census Annual Research Conference*, 571-581.

FINDLEY, D.F., MONSELL, B.C., OTTO, M.C., BELL, W.R., and PUGH, M.G. (1998). New capabilities and methods of the X-12 ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177.

FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1992). Estimates for longitudinal surveys with applications to the U.S. Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 301-324.

GRAY, A., and THOMSON, P. (1996). Design of moving-average trend filters using fidelity, smoothness and minimum revisions criteria. *Time Series Analysis in Memory of E.J. Hannan.* (Ed. P. Robinson and M. Rosenblatt), 205-219. Springer lecture notes in statistics, 115.

HENDERSON, R. (1916). Note on graduation by adjusted averages. *Transactions of the Actuarial Society of America*, 17, 43-48.

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

KALTON, G., and CITRO, C.F. (1993). Panels surveys: adding the fourth dimension. *Survey Methodology*, 19, 205-215.

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

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KNOWLES, J. (1997). Trend Estimation Practices of National Statistical Institutes. Office for National Statistics, Methods and Quality Division, UK, MQ 044.

KNOWLES, J., and KENNY (1997). An Investigation of Trend Estimation Methods. Office for National Statistics, Methods and Quality Division, UK, MQ 044.

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

LINACRE, S., and ZARB, J. (1991). Picking turning points in the economy. *Australian Economic Indicators*, Australian Bureau of Statistics, catalogue no. 1350.0.

MCLAREN, C.H. (1999). Designing Rotation Patterns and Filters for Trend Estimation in Repeated Surveys. Unpublished PhD Thesis, School of Mathematics and Applied Statistics, University of Wollongong.

MCLAREN, C.H., and STEEL, D.G. (1997). The effect of different rotation patterns on the sampling variance of seasonal and trend filters. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1997, 790-795.

MIAZAKI, E.S., and DOREA, C.C.Y. (1993). Estimation of the parameters of a time series subject to the error of rotation sampling. *Communications in Statistics*, A, 22, 805-825.

PFEFFERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116.

RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 69, 492-509.

SCOTT, A.J., SMITH T.M.F., and JONES, R. G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 3-73.

SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). *X-11 Variant of the Census Method II Seasonal Adjustment Program.* Technical Paper 15, Bureau of the Census, U.S. Department of Commerce, Washington, D.C.

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

SMITH, T.M.F. (1997). Discussion of paper by Steel. *Journal of the Royal Statistical Society A*, 160, 33-34.

STEEL, D.G. (1996). Options for Producing Monthly Estimates of Unemployment According to the ILO Definition. Central Statistical Office, U.K.

STEEL, D.G. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition. *Journal of the Royal Statistical Society A*, 160, 5-46.

STEEL, D.G., and MCLAREN, C.H. (2000). The effect of different rotation patterns on the revisions of trend estimates. *Journal of Official Statistics*, 16, 61-76.

STEEL, D.G., and DEMEL, R. (1988). The Contribution of Sampling Error to the Variability of Statistical Series. Paper Presented at the National Mathematical Sciences Congress, Canberra.

SUTCLIFFE, A. (1993). *X-11 Time Series Decomposition and Sampling Errors.* Working Papers in Econometrics and Applied Statistics, No 93/2. Australian Bureau of Statistics, catalogue no 1351.

SUTCLIFFE, A., and LEE, G. (1995). Seasonal Analysis and Sample Design. Paper presented at the Conference of Survey Measurement and Process Quality. Bristol 1995.

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.

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

WOLTER, K.M., and MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (D. Krewski, R. Platek and J.N.K. Rao, Eds.). New York: Academic Press, 367-407.

YOUNG, A.H. (1968). Linear Approximations to the Census and BLS Seasonal Adjustment Methods. *Journal of the American Statistical Association*, 63, 445-471.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 173-181
Statistics Canada

Hierarchical Bayes Estimation of Small Area Means Using Multi-Level Models

YONG YOU and J.N.K. RAO[1]

ABSTRACT

Standard multi-level models with random regression parameters are considered for small area estimation. We also extend the models by allowing unequal error variances or by assuming random effect models for both regression parameters and error variances. We present these models in a hierarchical Bayes framework and estimate a small area mean by its posterior mean. Posterior variance of the small area mean is used as a measure of precision of the estimate. It automatically takes into account the extra uncertainty associated with the hyperparameters in the multi-level model. Gibbs sampling is used to compute the posterior means and posterior variances of small area means. Rao-Blackwellized estimators that reduce the Monte Carlo errors are obtained. Bayesian model selection and sensitivity analysis are also studied. The procedure is illustrated using data on household income in some counties (small areas) of Brazil.

KEY WORDS: Gibbs sampling; Hierarchical Bayes; Multi-level model; Sampling error variance; Small area.

1. INTRODUCTION

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area estimators. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas; in particular, model-based indirect estimators. Battese, Harter and Fuller (1988) proposed and applied a nested error regression model to provide model-based small area estimates. The model takes the form

$$y_{ij} = x_{ij}^T \beta + v_{0i} + e_{ij}, j = 1, ..., n_i; i = 1, ..., m, \qquad (1)$$

where y_{ij} are the observations associated with the sampled units in the i-th small area, $i = 1, ..., m$, x_{ij} is the $p \times 1$ vector of unit-level explanatory variables, β is a set of p fixed regression parameters, v_{0i} are independent area effects with $E(v_{0i}) = 0$ and $V(v_{0i}) = \sigma_v^2$. The e_{ij}'s are assumed to be independent random error variables with $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_e^2$. v_{0i} and e_{ij} are also assumed to be independent. For the whole population, model (1) applies with n_i replaced by N_i, the small area population size. The model (1) may be expressed in matrix notation as follows

$$Y_i = X_i \beta + v_{0i} 1_{n_i} + e_i, i = 1, ..., m,$$

where $Y_i = (y_{i1}, ..., y_{i,n_i})^T$, $X_i = (x_{i1}, ..., x_{i,n_i})^T$ is a $n_i \times p$ matrix, $1_{n_i} = (1, ..., 1)^T$ is the unit vector of length n_i, and $e_i = (e_{i1}, ..., e_{i,n_i})^T$.

Holt and Moura (1993) extended the above framework to a multi-level model by introducing random regression coefficients and then relating them to area-level explanatory variables to explain some of the between small area variation. The model can be stated as follows:

$$Y_i = X_i \beta_i + e_i, \beta_i = Z_i \gamma + v_i \qquad (2)$$

where Z_i is the $p \times q$ design matrix of area-level variables, γ is a $q \times 1$ vector of fixed coefficients, and $v_i = (v_{i1}, ..., v_{ip})^T$ is a $p \times 1$ vector of random effects for the i-th area. The v_i's are independent across areas and have a joint distribution within each area with $E(v_i) = 0$ and $V(v_i) = \Phi$, where the variance covariance matrix Φ is unknown. Note that model (2) effectively integrates the use of unit-level and area-level covariates into a single model. Holt and Moura (1993) and Moura and Holt (1999) extended Prasad and Rao's (1990) framework to the above multi-level model to get the best linear unbiased predictor (BLUP) of the small area mean $\mu_i = \bar{X}_i^T \beta_i$ assuming that N_i is large, where $\bar{X}_i$ is the $p \times 1$ vector of known population means of the auxiliary variables for the i-th small area. They also obtained the empirical BLUP (EBLUP) and a second order approximation to the mean squared error (MSE) of EBLUP for the multi-level model. Using household income data in some counties (small areas) of Brazil, they demonstrated gain in efficiency of the EBLUP estimators over the EBLUP estimators obtained from nested error regression models. Ghosh and Rao (1994) and Rao (1999) provide a detailed overview of model-based methods for small area estimation.

In this paper, we study the multi-level model (2) in a hierarchical Bayes framework and extend the model to more general multi-level models which allow fixed unequal error variances or random error variances. The small area mean μ_i is estimated by its posterior mean and its precision is measured by its posterior variance. Posterior variance automatically takes into account the extra uncertainty

[1] Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

associated with the hyperparameters in the multi-level model. We use the Gibbs sampling method to compute the hierarchical Bayes estimates and the associated posterior variances. Section 2 presents the hierarchical Bayes multi-level models with different assumptions on error variances and related Gibbs sampling inference. Section 3 illustrates our methodology and studies model selection and sensitivity analysis by employing data on household incomes in some counties (small areas) of Brazil. And finally in section 4, we give some comments and concluding remarks.

2. MULTI-LEVEL MODELS AND GIBBS SAMPLING INFERENCE

2.1 Equal Error Variances

We consider a hierarchical Bayes representation of the multi-level model (2) as follows:

Model 1:

(i) Conditional on β_i and σ_e^2, y_{ij}'s are independent with

$$y_{ij}|\beta_i, \sigma_e^2 \sim N(x_{ij}^T\beta_i, \sigma_e^2), \quad (i = 1, ..., m; j = 1, ..., n_i); \tag{3}$$

(ii) Conditional on γ and Φ, β_i's are independent with

$$\beta_i|\gamma, \Phi \sim N_p(Z_i\gamma, \Phi), (i = 1, ..., m). \tag{4}$$

To complete our Bayesian model specification, we adopt the prior distributions for parameters as follows:

(iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_e \sim G(a, b)$ and $\Omega \sim W_p(\alpha, R)$, where $\tau_e = \sigma_e^{-2}$, $\Omega = \Phi^{-1}$, and D, a, b, α and R are known.

In step (iii) of Model 1, $G(a,b)$ denotes a gamma distribution with density given by $f(x) = b^a/\Gamma(a)x^{a-1}e^{-xb}$, $a>0, b>0, x\geq 0$, and $W_p(\alpha, R)$ is a Wishart distribution with density function

$$f(X) = \frac{|R|^{\frac{\alpha}{2}}}{2^{\alpha p/2}\Gamma_p\left(\frac{\alpha}{2}\right)}|X|^{\frac{\alpha-p-1}{2}}\exp\{-\frac{1}{2}\text{tr}(RX)\},$$

where $X>0, R>0$ and $\Gamma_p(\alpha)$ is multivariate gamma function defined as

$$\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(\alpha + \frac{1}{2}(1-j)\right).$$

Remark 1.1: The prior distributions in step (iii) are conjugate with the sampling and population distributions given by (3) and (4) in the sense that they lead to full conditional distributions for γ, τ_e and Ω that are again normal, gamma and Wishart distribution, respectively. The Wishart distribution is the multivariate version of gamma distribution for the inverse variance covariance matrix of random effects. The importance of conjugacy may be exploited as follows: (1) In the Gibbs sampling step, without conjugacy the full conditional distribution for any parameter will be known up to normalizing constants. In this case, more sophisticated random generation will be required. (2) Closed-form full conditional distributions may be employed to find the Rao-Blackwellized estimators of the posterior means and posterior variances, and thus to improve posterior estimation. In general, for Bayesian inference, choosing priors is not a simple job because any proper prior on the model parameters is a plausible candidate. This is a limitation of Bayesian methods.

Remark 1.2: It is important to note that we have used proper priors on all the unknown parameters to ensure that all the posterior distributions are proper (Hobert and Casella 1996). Hence we do not face the problem of some posteriors being improper. Values for the parameters of the priors (*i.e.*, hyperparameters) are chosen to reflect a fairly vague knowledge of the prior distributions. Details will be given in section 3 on data analysis.

Remark 1.3: In Model 1, we assume equal error variance σ_e^2 for all small areas. In practice, however, variances of sampling error could be different for different small areas. A more general model should allow possibly different error variances. In sections 2.2 and 2.3, we will introduce unequal error variance and random error variance models.

We are interested in finding the posterior distributions of β_i's given the data $Y = (\{y_{ij}\}, i = 1, ..., m; j = 1, ..., n_i)$, and in particular in finding the posterior estimates of small area means $\mu_i = \bar{X}_i^T\beta_i$, which depend on the estimates of β_i. Direct evaluation of the joint posterior distribution involves high-dimensional numerical integration, and is not computationally feasible. Therefore, we use the Gibbs sampling method (Gelfand and Smith 1990) to generate samples from the joint posterior distributions. To implement the Gibbs sampling under Model 1, we need the full conditional distributions given by:

(i) For $i = 1, ..., m$,

$$[\beta_i|Y, \gamma, \Omega, \tau] \overset{\text{ind}}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i\gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii) $$[\gamma|Y, \beta, \Omega, \tau_e] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T\Omega Z_i + D^{-1}\right)^{-1}\left(\sum_{i=1}^m Z_i^T\Omega\beta_i\right), \left(\sum_{i=1}^m Z_i^T\Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iii) $$[\Omega|Y, \beta, \gamma, \tau] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^m(\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(iv) $$[\tau_e|Y, \beta, \gamma, \Omega] \sim G\left(a + \frac{1}{2}\sum_{i=1}^m n_i, b + \frac{1}{2}\left(\sum_{i=1}^m (Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\right)\right).$$

Since all the full conditional distributions have closed-form, it is easy to generate samples. Gibbs sampling method is as follows: (a) Using starting values $\gamma^{(0)}$, $\Omega^{(0)}$ and $\tau_e^{(0)}$, draw $\beta_i^{(1)}$, $i = 1, ..., m$, from $[\beta_i | Y, \gamma, \Omega, \tau_e]$; (b) Draw $\gamma^{(1)}$ from $[\gamma | Y, \beta, \Omega, \tau_e]$ using $\beta_i^{(1)}$, $i = 1, ..., m$, $\Omega^{(0)}$ and $\tau_e^{(0)}$; (c) Draw $\Omega^{(1)}$ from $[\Omega | Y, \beta, \gamma, \tau_e]$ using $\beta_i^{(1)}$, $i = 1, ..., m$, $\gamma^{(1)}$ and $\tau_e^{(0)}$; (d) Draw $\tau_e^{(1)}$ from $[\tau_e | Y, \beta, \gamma, \Omega]$ using $\beta_i^{(1)}$, $i = 1, ..., m$, $\gamma^{(1)}$ and $\Omega^{(1)}$. Steps (a)-(d) complete one sampling cycle. Perform a large number of cycles, say t, called "burn-in" period, until convergence and then treat $\{\beta_i^{(t+k)}, i = 1, ..., m; \gamma^{(t+k)}; \Omega^{(t+k)}; \tau_e^{(t+k)}; k = 1, ..., G\}$ as G samples from the joint posterior of β_i, $i = 1, ..., m$, γ, Ω and τ_e.

Suppose a sample of size G is obtained as $\{\beta_i^{(k)}, i = 1, ..., m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, ..., G\}$. To obtain an estimator of the posterior mean of β_i, one can use the sample mean of the $\{\beta_i^{(k)}\}$. Since β_i has a closed form full conditional distribution, we can use the sample mean of the conditional expectations $\{E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$ to improve our estimation, since $E(\beta_i | Y) = E(E(\beta_i | Y, \gamma, \Omega, \tau_e))$, and $\mathrm{Var}(\beta_i | Y) \geq \mathrm{Var}(E(\beta_i | Y, \gamma, \Omega, \tau_e))$. This modification is based on the well-known Rao-Blackwell theorem and the corresponding estimator is the so-called Rao-Blackwellized estimator (Gelfand and Smith 1990, 1991). Thus we have the following two alternative estimators for β_i:

$$\hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^{G} \beta_i^{(k)} \tag{5}$$

and

$$\hat{\beta}_i^{(\mathrm{RB})} = \frac{1}{G} \sum_{k=1}^{G} E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})$$

$$= \frac{1}{G} \sum_{k=1}^{G} (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1}$$

$$(\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}), \tag{6}$$

where $\hat{\beta}_i^{(E)}$ is the empirical estimator and $\hat{\beta}_i^{(\mathrm{RB})}$ is the Rao-Blackwellized estimator. Both $\hat{\beta}_i^{(E)}$ and $\hat{\beta}_i^{(\mathrm{RB})}$ are unbiased for the posterior mean. However, $\hat{\beta}_i^{(\mathrm{RB})}$ is better than $\hat{\beta}_i^{(E)}$ in terms of simulation standard error (Gelfand and Smith 1991).

The corresponding estimators for the small area mean μ_i are given as

$$\hat{\mu}_i^{(E)} = \bar{X}_i^T \hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^{G} \bar{X}_i^T \beta_i^{(k)} \tag{7}$$

and

$$\hat{\mu}_i^{(\mathrm{RB})} = \bar{X}_i^T \hat{\beta}_i^{(\mathrm{RB})} = \frac{1}{G} \sum_{k=1}^{G} \bar{X}_i^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1}$$

$$(\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}). \tag{8}$$

We anticipate that both $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(\mathrm{RB})}$ will give almost the same point estimates. However, it will be of interest to compute and compare the simulation standard errors of these two estimators to evaluate the effects of Rao-Blackwellization; see section 3.

To obtain the posterior variance of μ_i, we first find the posterior variance of β_i, since $V(\mu_i | Y) = \bar{X}_i^T V(\beta_i | Y) \bar{X}_i$. Note that

$$V(\beta_i | Y) = E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + V(E(\beta_i | Y, \gamma, \Omega, \tau_e))$$
$$= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + E(E(\beta_i | Y, \gamma, \Omega, \tau_e)^2)$$
$$- [E(E(\beta_i | Y, \gamma, \Omega, \tau_e))]^2. \tag{9}$$

Using (9), the Rao-Blackwellized estimator of the posterior variance of β_i, denoted by $\hat{V}(\beta_i)$, can be obtained using the Gibbs samples $\{\beta_i^{(k)}, i = 1, ..., m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, ..., G\}$; see Appendix A1. The posterior variance of small area mean μ_i is then estimated by

$$\hat{V}(\mu_i) = \bar{X}_i^T \hat{V}(\beta_i) \bar{X}_i. \tag{10}$$

The same estimation procedure can be applied to the sampling error variance σ_e^2. Since conditionally σ_e^2 has an inverse gamma distribution, the Rao-Blackwellized estimator of the posterior mean of σ_e^2 is obtained as

$$\hat{\sigma}_e^{2(\mathrm{RB})} = \frac{1}{G} \sum_{k=1}^{G} \left[b + \frac{1}{2} \sum_{i=1}^{m} (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)}) \right]$$

$$\left(a + \frac{1}{2} \sum_{i=1}^{m} n_i - 1 \right)^{-1}. \tag{11}$$

Since we are mainly interested in estimating the small area means, calculation of the sampling variance is only for the purpose of model selection. Details on model selection will be given in section 3.2.

2.2 Unequal Error Variances

In practice, it is more realistic to allow unequal error variances for the sampling errors. Let σ_i^2 be the true sampling error variance for the i-th small area. A straightforward extension of Model 1 leads to the following hierarchical Bayes multi-level unequal error variance model:

Model 2:

(i) Conditional on β_i and σ_i^2, y_{ij}'s are independent with

$$y_{ij} | \beta_i, \sigma_i^2 \sim N(x_{ij}^T \beta_i, \sigma_i^2),$$
$$(i = 1, ..., m; j = 1, ..., n_i); \tag{12}$$

(ii) Conditional on γ and Φ, β_i's are independent with

$$\beta_i | \gamma, \Phi \sim N_p(Z_i \gamma, \Phi), (i = 1, ..., m); \tag{13}$$

(iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_i \overset{ind}{\sim} G(a_i, b_i)$, and $\Omega \sim W_p(\alpha, R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$, and D, a_i, b_i, α and R are known.

Remark 2.1: Model 2 reduces to Model 1 when $\sigma_i^2 = \sigma_e^2$ for all i. From a hierarchical Bayes perspective, extension from the equal error variance model to the unequal error variance model is straightforward. Also there is no difficulty in the Gibbs sampling implementation.

Remark 2.2: τ_i's are assumed to be independent and have prior distributions $G(a_i, b_i)$, where a_i and b_i are known hyperparameters and usually chosen to be very small to reflect a vague knowledge about τ_i's.

The full conditional distributions for Gibbs sampling under Model 2 are given by:

(i) For $i = 1, ..., m$,
$$[\beta_i | Y, \gamma, \Omega, \tau] \overset{ind}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1} (\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii)
$$[\gamma | Y, \beta, \Omega, \tau] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i\right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iii)
$$[\Omega | Y, \beta, \gamma, \tau] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^m (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(iv) For $i = 1, ..., m$,
$$[\tau_i | Y, \beta, \gamma, \Omega] \sim G(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2} \times (Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)).$$

For Model 2, the empirical estimators of the posterior means of β_i and μ_i have the same form as (5) and (7). The Rao-Blackwellized estimators $\hat{\beta}_i^{(RB)}$ and $\hat{\mu}_i^{(RB)}$ are the estimators given by (6) and (8) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$. Estimator of posterior variance is $\hat{V}(\mu_i)$ given by (10) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$.

For the purpose of model selection and model comparison, we also find the Rao-Blackwellized estimator of the posterior mean of σ_i^2 under Model 2 as

$$\hat{\sigma}_i^{2(RB)} = \frac{1}{G}\sum_{k=1}^G \left[b_i + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T (Y_i - X_i\beta_i^{(k)})\right] \times (a_i + \frac{1}{2}n_i - 1)^{-1}. \quad (14)$$

2.3 Random Error Variances

In Model 2, we assumed unequal error variances for the sampling errors. Kleffe and Rao (1992) used a simple random error variance model to derive the best linear unbiased predictors for small area means. In this section we extend their model to the multi-level case. We assume random effect models on both regression coefficients β_i and sampling error variances σ_i^2, which leads to Model 3 given below.

Model 3:

(i) Same as in Model 2;

(ii) Same as in Model 2;

(iii) Conditional on η and λ, τ_i's are independent with
$$\tau_i | \eta, \lambda \overset{iid}{\sim} G(\eta, \lambda), \quad (15)$$
where $\tau_i = \sigma_i^{-2}$;

(iv) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\Omega \sim W_p(\alpha, R)$, $\eta \sim U^+$ and $\lambda \sim U^+$, where U^+ denotes a uniform distribution over a subset of R^+ with large but finite length, D, α and R are known.

Remark 3.1: In Model 3, we assume that τ_i's are iid gamma random variables with unknown hyperparameters η and λ. Thus we have population models for both regression coefficient β_i and sampling variance σ_i^2. In Model 1 and Model 2, we considered modelling β_i only and assumed vague proper prior distributions on σ_e^2 or σ_i^2.

Remark 3.2: Assumption (iii) may not be a good population model for all τ_i's. Alternatively, we can model τ_i in a more realistic way, as in the case of β_i, by specifying a regression model for the logarithm of τ_i. This may require some auxiliary information related to τ_i. In the data analysis of section 3, however, we simply used $G(\eta, \lambda)$ as the population model for τ_i. Generally it is not easy to model the sampling variances when they are unknown.

The full conditional distributions for Gibbs sampling under Model 3 are given by:

(i) For $i = 1, ..., m$,
$$[\beta_i | Y, \tau, \gamma, \Omega, \eta, \lambda] \overset{ind}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1} (\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii) For $i = 1, ..., m$,
$$[\tau_i | Y, \beta, \gamma, \Omega, \eta, \lambda] \overset{ind}{\sim} G\left(\eta + \frac{n_i}{2}, \frac{1}{2}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i) + \lambda\right)$$

(iii)
$$[\gamma | Y, \beta, \tau, \eta, \lambda] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i\right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iv)
$$[\Omega | Y, \beta, \sigma^2, \gamma, \eta, \lambda] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^m (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(v) $$[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda] \propto [\Gamma(\eta)]^{-m}\lambda^{m\eta}(\Pi_{i=1}^{m}\tau_i)^{\eta}$$

(vi) $$[\lambda|Y,\beta,\tau,\gamma,\Omega,\eta] \sim G\left(m\eta+1, \textstyle\sum_{i=1}^{m}\tau_i\right)$$

For Model 3, the posterior estimators of β_i and μ_i have the same forms as those given for Model 2. Under Model 3, the Rao-Blackwellized estimator of the posterior mean of σ_i^2 is given by

$$\hat{\sigma}_i^{2(RB)} = \frac{1}{G}\sum_{k=1}^{G}[\lambda^{(k)} + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T \times (Y_i - X_i\beta_i^{(k)})](\eta^{(k)} + \frac{1}{2}n_i - 1)^{-1}. \quad (16)$$

Under Model 3, $[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$ is known only up to a multiplicative constant. However, since $[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$ is a log-concave function of η (see Appendix A2), adaptive rejection sampling method of Gilks, Best and Tan (1995) can be used in the Gibbs sampler to generate samples from the conditional distribution $[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$.

3. DATA ANALYSIS

3.1 Data and Model Description

Following Holt and Moura (1993) and Moura and Holt (1999), we considered the estimation of household income in some counties (small areas) of Brazil. Holt and Moura's original data contains 140 small areas with the sampling units taken from each area by simple random sampling. The hierarchical Bayes method does not require the number of small areas to be large, unlike in the case of EBLUP method, for getting standard errors. Therefore, we used only a small part of the original data set in our data analysis for simple illustration. Our data set contains a subset of 10 small areas with 28 sampling units obtained by simple random sampling in each area.

Let y_{ij} denote the j-th household's income in the i-th small area. There are two unit level auxiliary variables, namely x_1 and x_2, where x_1 denotes the number of rooms in a household and x_2 denotes the educational attainment of Head of Household. The sampling model is given by

$$y_{ij} = x_{ij}^T\beta_i + e_{ij} = \beta_{0i} + x_{1ij}\beta_{1i} + x_{2ij}\beta_{2i} + e_{ij}, \quad (17)$$

where x_{1ij} denotes the number of rooms in the j-th household of small area i and x_{2ij} denotes the corresponding educational attainment of Head of Household. Values of x_{1ij} and x_{2ij} are centered around their respective overall sample means and e_{ij} is the sampling error variable with its distribution specified by the three error variance models discussed in section 2.

In the sampling model (17), β_i is the random regression coefficient corresponding to the i-th small area and is modelled as

$$\beta_{0i} = \gamma_0 + \nu_{0i}, \beta_{1i} = \gamma_{10} + \gamma_{11}z_i + \nu_{1i}, \beta_{2i} = \gamma_{20} + \gamma_{21}z_i + \nu_{2i},$$

where $\gamma = (\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})^T$ is the unknown vector of fixed regression parameters, $\nu_i = (\nu_{0i}, \nu_{1i}, \nu_{2i})^T$ is the i-th small area random effect vector distributed as $\nu_i \sim N_3(0, \Phi)$, and z_i is an area level variable defined as the average number of cars per household in each small area. Value of z_i is also centered around its overall sample mean.

We used the three models discussed in section 2 for our data analysis. Vague proper prior distributions on unknown parameters are specified as follows: $\gamma \sim N_5(0, D)$ where $D = \text{diag}(10^4, 10^4, 10^4, 10^4, 10^4)$, thus $\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}$ are assumed to be independent normal variables with a mean of 0 and a standard deviation of 100, so that a 95% prior interval is around ±200, and the prior will be locally uniform over the region supported by the likelihood. Alternatively a uniform prior on a suitably wide interval could be given, such as U(−200, 200). A Wishart prior $W_3(\alpha, R)$ is specified for the inverse covariance matrix $\Omega = \Phi^{-1}$. To represent vague prior knowledge, we have chosen the degrees of freedom α for this distribution to be as small as possible, *i.e.*, $\alpha = 3$, the rank of Ω(Spiegelhalter, Thomas, Best and Gilks 1996). The scale matrix R is specified with diagonal elements equal to 1 and off-diagonal elements equal to 0.001, which represents our prior guess at the order of magnitude of the covariance matrix. For Model 1 and Model 2, a gamma prior G(0.001, 0.001) is assumed for τ_e and τ_i's. For Model 3, $\tau_i \sim G(\eta, \lambda)$, and η and λ are assumed to be independently distributed as U(0,10000), *i.e.*, the uniform distribution over a large interval. We anticipate that the vague proper priors on the hyperparameters would approximate the flat priors reasonably well and thus would have minimal effect on the posterior estimation.

We implemented the Gibbs sampler for the three models using the BUGS program (Spiegelhalter *et al.* 1996), aided by CODA Splus function (Best, Cowles and Vines 1996) for assessing convergence. The BUGS program constructs the necessary full conditional distributions and carries out the Gibbs sampling as long as we specify our models using the BUGS language. Priors and initial values of the parameters must be specified in the program. For each model, the Gibbs sampler was first run for a "burn-in" period of 2,000 iterations, then 5,000 more iterations were run and kept for model analysis and estimation.

Our interest is to estimate the small area mean $\mu_i = \bar{X}_i^T\beta_i = \beta_{0i} + \bar{X}_{1i}\beta_{1i} + \bar{X}_{2i}\beta_{2i}$, where $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are the i-th small area population means of the auxiliary variables x_1 and x_2, respectively. For this, we will first select a model for the data set, then we will present the model-based estimates for the small area means based on the selected model.

3.2 Model Selection

We have proposed three models in section 2 based on different assumptions on sampling variances. To examine

which model fits the data, we first obtained the posterior estimates of the sampling variances under the three models. We also calculated the ordinary least square (OLS) estimates of the sampling variances within each area using only the area-specific data. Table 1 shows the Rao-Blackwellized estimates of the sampling variances under the three models as well as the OLS estimates.

Table 1
Estimated Sampling Error Variances

Area	OLS	Model 1	Model 2	Model 3
1	38.17	76.86	40.18	63.60
2	31.75	76.86	34.24	62.13
3	81.26	76.86	94.77	79.58
4	48.73	76.86	52.01	67.27
5	115.98	76.86	121.65	87.70
6	90.74	76.86	94.35	79.78
7	101.67	76.86	101.67	82.14
8	135.65	76.86	159.94	97.96
9	59.10	76.86	63.37	70.57
10	62.86	76.86	65.72	71.22

From Table 1, the OLS estimates indicate large variations among the ten small areas. Model 1 assumes an equal error variance σ_e^2 for all areas and σ_e^2 is estimated by $\hat{\sigma}_e^{2(RB)} = 76.86$, which is much smaller than the OLS estimates for some areas. Model 2 assumes unequal error variance σ_i^2 across areas. Under Model 2, the estimated error variances $\hat{\sigma}_i^{2(RB)}$ to some extent show the feature of the areas; $\hat{\sigma}_i^{2(RB)}$ are consistent with the pattern of the OLS estimates. The most notable result is $\hat{\sigma}_5^{2(RB)} = 121.65$ and $\hat{\sigma}_8^{2(RB)} = 159.94$, which show that there are larger variations within small areas 5 and 8. Model 3 assumes σ_i^2's to be random variables distributed as $G(\eta, \lambda)$. Under Model 3, all $\hat{\sigma}_i^{2(RB)}$ tend to be equal to and have moved toward $\hat{\sigma}_e^{2(RB)} = 76.86$. The results in Table 1 suggest that Model 2, the unequal error variance model, could be the best model for our data set. For further investigation, we now present a cross-validation study to select a best fit model.

In order to study how the data support each model, we calculated the cross-validation predictive densities for each data point y_{ij}. The cross-validation density for y_{ij} is the conditional density $f(y_{ij}|Y_{(ij)})$, where $Y_{(ij)}$ denotes all data except y_{ij}. We looked at the value of $f(y_{ij}|Y_{(ij)})$ at the observed data point, the so called conditional predictive ordinate, or CPO, for each of the three models. That is

$$\text{CPO}_{ij} = f(y_{ij,\,\text{obs}}|Y_{(ij),\,\text{obs}}),$$

where $y_{ij,\,\text{obs}}$ denotes the observed data point. Since CPOs are nothing but the observed likelihoods, models with larger CPOs provide better fit to the observed data. By using the output from the Gibbs sampler, we can calculate the CPOs for all data points. For example, under Model 1, we have

$$f(y_{ij}|Y_{(ij)}) = \frac{f(Y)}{f(Y_{(ij)})} = \frac{1}{\int \frac{f(Y_{(ij)}, \beta_i, \sigma_e^2)}{f(Y, \beta_i, \sigma_e^2)} \cdot f(\beta_i, \sigma_e^2|Y) d\beta_i d\sigma_e^2} = \frac{1}{\int \frac{1}{f(y_{ij}|Y_{(ij)}, \beta_i, \sigma_e^2)} \cdot f(\beta_i, \sigma_e^2|Y) d\beta_i d\sigma_e^2}.$$

Now noting that the y_{ij}'s are conditionally independent, *i.e.*, $f(y_{ij}|Y_{(ij)}, \beta_i, \sigma_e^2) = f(y_{ij}|\beta_i, \sigma_e^2)$, the CPO values are calculated as

$$\widehat{\text{CPO}}_{ij} = \frac{1}{\frac{1}{G}\sum_{k=1}^{G} \frac{1}{f(y_{ij,\,\text{obs}}|\beta_i^{(k)}, \sigma_e^{2(k)})}}, \tag{18}$$

where $f(y_{ij}|\beta_i, \sigma_e^2)$ is the normal density function given by (3). For Model 2 and Model 3, the CPOs are calculated with $\sigma_e^{2(k)}$ replaced by $\sigma_i^{2(k)}$ in (18). More detailed discussion can be found in Gelfand (1995).

We present a CPO plot for the three models in Figure 1. Clearly Model 2 is the best model among the three, because a majority of CPO values for Model 2 are significantly larger than those for Model 1 and Model 3. Model 3 is slightly better than Model 1 in terms of CPO values. Also there are small CPO values for all three models, which indicate that our model assumptions may not be very well satisfied by our data set.

According to the sampling variance estimates given in the Table 1 and the CPO plot, we conclude that Model 2 is a good model for our data. Therefore, we used Model 2 to find model-based estimates of small area means and associated posterior standard errors.

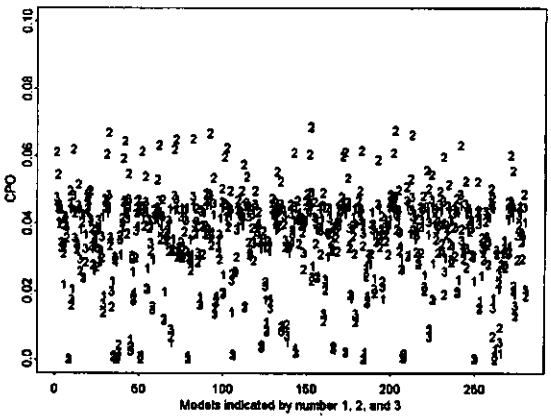


Figure 1. Model selection: CPO comparison plot

3.3 Result of Estimation

We now present the estimates of the small area means based on Model 2 only. Table 2 presents the estimated posterior small area means and the corresponding posterior

standard errors. Our study found that the empirical estimator $\hat{\mu}_i^{(E)}$ and the Rao-Blackwellized estimator $\hat{\mu}_i^{(RB)}$ gave almost the same point estimates, thus we only reported the estimates obtained by using $\hat{\mu}_i^{(RB)}$. For comparison, we also calculated the direct estimates (sample means) and corresponding direct standard errors for the ten areas. It is clear from Table 2 that the model-based estimates are substantially more efficient than the direct estimates. The posterior standard errors are much smaller than the direct standard errors.

Table 2
Estimates of Small Area Means

Area	$\bar{y}_i$	s.e.	$\hat{\mu}_i^{(RB)}$	s.e.
1	11.08	9.53	10.23	0.81
2	7.91	6.82	9.84	0.85
3	13.48	14.15	13.01	1.08
4	6.53	8.01	10.95	1.11
5	19.52	14.96	17.87	1.57
6	11.21	11.38	10.21	0.93
7	8.72	11.24	9.58	0.97
8	12.81	13.99	10.30	1.19
9	10.18	8.76	11.34	1.01
10	10.01	11.30	9.79	0.87

In order to study the effects of Rao-Blackwellization, we calculated the simulation standard errors of $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$, which are respectively the sample standard errors of $\{\bar{X}_i^T\hat{\beta}_i^{(k)}\}$ and $\{\bar{X}_i^T E[\beta_i|Y,\gamma^{(k)},\Omega^{(k)},\tau_e^{(k)}]\}$. Table 3 presents the simulation standard errors. It is clear from Table 3 that the Rao-Blackwellized estimator $\hat{\mu}_i^{(RB)}$ has much smaller simulation standard error than the empirical estimator $\hat{\mu}_i^{(E)}$ for all areas. In all cases the standard error of $\hat{\mu}_i^{(RB)}$ is about 50% to 75% of the standard error of $\hat{\mu}_i^{(E)}$, demonstrating the benefit of Rao-Blackwellization. Thus $\hat{\mu}_i^{(RB)}$ is more stable than $\hat{\mu}_i^{(E)}$ when used to produce point estimates for the posterior means in computational Bayesian analysis. It should be mentioned that the simulation standard error of $\hat{\mu}_i^{(E)}$ is also an estimator of the posterior standard error. Thus the simulation standard error of $\hat{\mu}_i^{(E)}$ in Table 3 is almost identical to the estimated standard error of $\hat{\mu}_i^{(RB)}$ in Table 2.

Table 3
Simulation Standard Errors

Area	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$
1	0.817	0.506
2	0.862	0.498
3	1.090	0.548
4	1.101	0.604
5	1.583	0.878
6	0.930	0.481
7	0.978	0.480
8	1.208	0.842
9	0.997	0.524
10	0.869	0.513

3.4 Sensitivity Analysis

In Model 2, the error variances $\tau_i = \sigma_i^{-2}$ are assumed to be independent with prior distributions $G(a_i, b_i)$ or σ_i^2 with the inverse gamma $IG(a_i, b_i)$, where a_i and b_i are known values chosen to reflect our prior knowledge about σ_i^2. In practice, it is always difficult to obtain accurate information about the sampling variances. Also, as the number of small areas m increases, the number of variance components σ_i^2 will increase. We are interested in the possible effects caused by the choice of priors on σ_i^2's; in particular, we would like to evaluate the sensitivity of the posterior means to the choice of priors on the sampling variances σ_i^2. In our data analysis, a_i and b_i were chosen to be 0.001. Thus we used proper priors with very small parameter values for the variance components to reflect our vague knowledge about σ_i^2. In order to test the sensitivity of the posterior estimates to the choice of a_i and b_i under Model 2, we set $a_i = b_i$ at six different values, *i.e.*, 0.0001, 0.001, 0.01, 0.1, 1, and 10. Since

$$\{\tau_i|Y,\beta,\gamma,\Omega\} \sim G\left(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)\right), \quad (19)$$

the sample effects $n_i/2$ and $(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)/2$ dominate the prior information a_i and b_i when a_i and b_i are small. Thus IG(0.0001, 0.0001), IG(0.001, 0.001), and IG(0.01,0.01) may be viewed as noninformative priors whereas IG(1, 1) and IG(10, 10) may be regarded as informative priors. Table 4 presents posterior means under Model 2 using the different priors on σ_i^2, and Table 5 presents the corresponding posterior variances.

Table 4
Comparison of Estimated Small Area Means

Small Area	$IG(a_i, b_i), a_i = b_i$					
	0	0	0.01	0.1	1	10
1	10.23	10.23	10.23	10.24	10.25	10.37
2	9.84	9.84	9.84	9.83	9.82	9.62
3	13.00	13.00	13.01	13.01	13.07	13.09
4	10.95	10.95	10.95	10.95	10.94	10.61
5	17.86	17.87	17.85	17.76	17.78	18.27
6	10.21	10.21	10.21	10.21	10.25	10.28
7	9.58	9.58	9.59	9.58	9.63	9.57
8	10.29	10.30	10.30	10.26	10.37	10.86
9	11.34	11.34	11.35	11.32	11.32	11.23
10	9.79	9.79	9.80	9.79	9.82	9.92

It is clear from Table 4 that the small area mean estimates are very stable: they are not sensitive to the choice of a_i and b_i. However, as shown in Table 5, the posterior variances decrease as the priors on σ_i^2 become more informative, and lead to smaller coefficients of variation (CV). This indicates that we can improve estimation results for small areas in terms of CV if we have more prior information on the sampling error variances. In our study, we only considered the case $a_i = b_i$. A more extensive study would involve different combinations of a_i and b_i.

Table 5
Comparison of Estimated Posterior Variances

Small Area	IG(a_i, b_i), $a_i = b_i$					
	0	0	0.01	0.1	1	10
1	0.658	0.658	0.658	0.656	0.653	0.499
2	0.724	0.724	0.724	0.711	0.684	0.462
3	1.167	1.167	1.167	1.161	1.152	0.917
4	1.220	1.220	1.218	1.217	1.202	0.919
5	2.455	2.455	2.454	2.462	2.139	1.335
6	0.871	0.870	0.870	0.830	0.826	0.699
7	0.933	0.933	0.931	0.930	0.914	0.779
8	1.418	1.417	1.418	1.375	1.351	1.337
9	1.015	1.014	1.014	1.011	0.975	0.790
10	0.760	0.760	0.760	0.750	0.745	0.613

Table 6 presents the posterior estimates of σ_i^2 using the different priors on σ_i^2. As we can see from Table 6, when a_i and b_i are small (< = 0.01), there is almost no difference among the estimates at all. As a_i and b_i increase, the estimates $\hat{\sigma}_i^{2(RB)}$ become smaller. However, if there is strong prior information on a_i and b_i, for example, $a_i = b_i = 10$, then the posterior estimates of σ_i^2 will be significantly different from the ones under noninformative priors.

Table 6
Comparison of Estimated Sampling Error Variances

Small Area	IG(a_i, b_i), $a_i = b_i$					
	0	0	0.01	0.1	1	10
1	40.09	40.1	40.05	39.64	37.14	22.29
2	34.19	34.18	34.17	33.97	31.74	19.05
3	94.48	94.49	94.42	93.76	86.73	50.60
4	52.08	52.08	52.04	51.63	48.21	28.82
5	121.60	121.70	121.60	121.40	113.70	66.75
6	94.03	94.03	93.83	92.96	87.21	52.90
7	102.30	102.30	102.20	101.40	94.85	57.58
8	160.10	160.00	159.90	159.10	147.60	86.61
9	63.46	63.46	63.38	62.99	58.46	34.85
10	65.88	65.87	65.89	65.40	60.76	36.60

4. CONCLUDING REMARKS

In this paper, we have presented hierarchical Bayes methods for small area estimation, using multi-level models. Clearly it is not easy to provide a suitable model for all small areas with satisfactory results, even if the Markov Chain Monte Carlo (MCMC) Bayesian methods such as the Gibbs sampling enable us to fit the data using Bayesian models of virtually unlimited complexity. The size and homogeneity of the areas and the availability of good auxiliary information will affect the final results. Models which prove suitable in some situations may be unsuitable in others. The hierarchical Bayes method also has some limitations such as the choice of priors on the model parameters and some sampling issues related to the Gibbs sampling method. Nevertheless, the general hierarchical Bayes methodology is applicable to a wide variety of situations for estimation of small area parameters. Model selection and choice is an important part of the hierarchical Bayes analysis. It is also important to compare the hierarchical Bayes method with other widely used methods in small area estimation, such as empirical Bayes (EB) and empirical best linear unbiased prediction (EBLUP). Work is in progress on extending our work to account for survey design weights, along the lines of You and Rao (1999).

ACKNOWLEDGMENTS

We would like to thank two referees and the Editor for their helpful comments and suggestions. We also would like to thank Professor F. Moura of Federal University of Rio de Janeiro in Brazil for providing the data set used in section 3. This work was partially supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

A1:

The Rao-Blackwellized estimator of the posterior variance of β_i is given by:

$$\hat{V}(\beta_i) = \frac{1}{G}\sum_{k=1}^{G} V(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})$$

$$+ \frac{1}{G}\sum_{k=1}^{G} [E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})]^2$$

$$- \left[\frac{1}{G}\sum_{k=1}^{G} E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})\right]^2$$

$$= \frac{1}{G}\sum_{k=1}^{G} (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1}$$

$$+ \frac{1}{G}\sum_{k=1}^{G} (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})$$

$$\times (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1}$$

$$- \frac{1}{G^2}\left[\sum_{k=1}^{G} (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})\right]$$

$$\times \left[\sum_{k=1}^{G} (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})\right]^T.$$

A2:

Lemma: $[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$ is a log-concave function of η.
Proof: Let $h(\eta)=\log[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$. It is enough to show that

$$\frac{\partial^2 h(\eta)}{\partial^2\eta}\leq 0.$$

Clearly,

$$\frac{\partial h(\eta)}{\partial\eta}=-m\frac{\Gamma'(\eta)}{\Gamma(\eta)}+m\log(\lambda)+\log(\Pi_{i=1}^{m}\tau_i).$$

Let $\psi(\eta)=\Gamma'(\eta)/\Gamma(\eta)$, then we have

$$\frac{\partial^2 h(\eta)}{\partial^2\eta}=-m\psi'(\eta)\leq 0$$

since $m>0$ and $\psi'(\eta)$ is positive on $(0,\infty)$(Temme, 1994, 54-55).

REFERENCES

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

BEST, N., COWLES, M.K., and VINES, K. (1996). CODA, *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*, Version 0.30. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

GELFAND, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, Eds.), 145-161. London: Chapman and Hall.

GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GELFAND, A.E., and SMITH, A.F.M. (1991). Gibbs sampling for marginal posterior expections. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

GILKS, W.R., BEST, N.G., and TAN, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics*, 44, 455-472.

HOBERT, J.P., and CASSELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.

HOLT, D., and MOURA, F. (1993). Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-30.

KLEFFE, J., and RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

MOURA, F., and HOLT, D. (1999). Small area estimation using multi-level models. *Survey Methodology*, 25, 73-80.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

SPIEGELHALTER, D., THOMAS, A., BEST, N., and GILKS, W. (1996). BUGS *0.5, Bayesian Inference Using Gibbs Sampling Manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

TEMME, N.M. (1994). *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. New York: John Wiley.

YOU, Y., and RAO, J.N.K. (1999). Pseudo hierarchical Bayes small area estimation using sampling weights. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 117-122.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 183-188
Statistics Canada

Double Sampling for Ratio and Regression Estimation With Sub-sampling the Non-respondents

FABIAN C. OKAFOR and HYUNSHIK LEE[1]

ABSTRACT

Cochran (1977, p. 374) proposed some ratio and regression estimators of the population mean using the Hansen and Hurwitz (1946) procedure of sub-sampling the non-respondents assuming that the population mean of the auxiliary character is known. For the case where the population mean of the auxiliary character is not known in advance, some double (two-phase) sampling ratio and regression estimators are presented in this article. The relative performances of the proposed estimators are compared with the estimator proposed by Hansen and Hurwitz (1946).

KEY WORDS: Hansen and Hurwitz estimator; Survey cost; Optimum sampling fraction.

1. INTRODUCTION

In many human surveys, information is in most cases not obtained from all the units in the survey even after some call-backs. An estimate obtained from such incomplete data may be misleading especially when the respondents differ from the non-respondents because the estimate can be biased. Hansen and Hurwitz (1946) proposed a technique for adjusting for non-response to address the bias problem. Their idea is to take a sub-sample from the non-respondents to get an estimate for the subpopulation represented by the non-respondents.

Cochran (1977), using Hansen and Hurwitz (1946) procedure, proposed the ratio and regression estimators of the population mean of the study variable in which information on the auxiliary variable is obtained from all the sample units, while some sample units failed to supply information on the study variable. In addition, the population mean of the auxiliary variable is known. In this paper we shall assume that the population mean of the auxiliary variable is not known. We, therefore, use the double sampling method to estimate the mean of the auxiliary variable and then go on to estimate the mean of the study variable in a similar manner as Cochran (1977).

In practice, non-response is often compensated for by weighting adjustment (Oh and Scheuren 1983) or by imputation (Kalton and Karsprzyk 1986). The procedures used for weighting adjustment and imputation strive for elimination of the bias due to non-response. However, those procedures are based on untenable assumptions on the response mechanism. When the assumed mechanism is wrong, then the resulting estimate can be seriously biased. Moreover, it is difficult to eliminate the bias entirely when non-response is confounded in the sense that the response probability is dependent on the survey character. Rancourt, Lee, and Särndal (1994) provided a partial correction for the situation. Hansen and Hurwitz's sub-sampling approach does not have this defect although it costs more because of extra work required for sub-sampling the non-respondents. Nonetheless, if the bias problem is serious, the procedure is a viable option to address the problem without resorting to 100 percent response, which can be very expensive.

In the next section, double sampling ratio and regression estimators are considered. Generally, the double sampling procedure is used when it is necessary to make use of auxiliary information to improve the precision of an estimate but the population distribution of the auxiliary information is not known. The first phase sample is used to estimate the population distribution of the auxiliary variable, while the second phase sample is used to obtain the required information on the variable of main interest. The optimum sampling fractions are derived for the estimators for a fixed cost. The performances of the proposed estimators are compared both theoretically and empirically with the Hansen and Hurwitz estimator.

2. THE DOUBLE SAMPLING RATIO AND REGRESSION ESTIMATORS

2.1 Background

To estimate the population mean $\bar{X}$ of the auxiliary variable, a large first phase sample of size n' is selected from N units in the population by simple random sampling without replacement (SRSWOR). A smaller second phase sample of size n is selected from n' by SRSWOR and the character y is measured on it. The ratio estimator of the mean of y is $\bar{y}'_r = (\bar{y}/\bar{x})\bar{x}'$, where $\bar{x}'$ is the sample mean from n' units, $\bar{y}$ and $\bar{x}$ are obtained from the second phase sample if there is no non-response in the second phase sample. If, however, there is non-response in the second phase sample, we may use an estimator obtained from only

[1] Fabian C. Okafor, Dept. of Statistics, University of Nigeria, Nsukka, Nigeria; Hyunshik Lee, formerly Statistics Canada, now Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

the respondents or take a sub-sample of the non-respondents and re-contact them. The former option is much cheaper than the latter because securing missing information from the non-respondents by re-contact requires usually much more effort and cost. However, it is quite feasible that the non-respondents differ significantly in the main character from the respondents so that a serious bias results. In this situation, sub-sampling of the non-respondents may be beneficial. Hence, we pursue the sub-sampling idea of Hansen and Hurwitz for a double sampling situation. Basically, the estimators proposed here are double sampling version of Cochran (1977, p. 374), that is, double sampling ratio and regression estimators for $\bar{Y}$ adjusted for non-response by using the Hansen and Hurwitz (1946) procedure.

Let's assume that all the n' units supplied information on the auxiliary variable x at the first phase. But let n_1 units supply information on y and n_2 refuse to respond at the second phase. From the n_2 non-respondents, an SRSWOR of m units is selected with the inverse sampling rate k, where $m = n_2/k$, $k > 1$. All the m units respond this time around. This can be applied in a household survey where the household size is used as an auxiliary variable for the estimation of, say, family expenditure. Information can be obtained completely on the family size during the household listing while there may be non-response on the household expenditure.

In the following presentation, we assume that the whole population (denoted by A) is stratified into two strata: one is the stratum (denoted by A_1) of N_1 units, which would respond on the first call at the second phase and the other stratum (denoted by A_2) consists of N_2 units, which would not respond on the first call at the second phase but will respond on the second call. Let the first and second phase samples be denoted by a' and a respectively, and let $a_1 = a \cap A_1$ and $a_2 = a \cap A_2$. The sub-sample of a_2 will be denoted by a_{2m}. Summation over the units in a set s will be denoted by Σ_s.

As a general rule, population parameters are denoted by capital letters except for Greek letters and the sample statistics by corresponding small letters.

2.2 The Double Sampling Ratio Estimator

We define the double sampling ratio estimator as follows:

$$d^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' = r^*\bar{x}' \tag{2.1}$$

where $\bar{x}^*$ and $\bar{y}^*$ are the Hansen-Hurwitz estimators for $\bar{X}$ and $\bar{Y}$, respectively, and are given by

$$\bar{u}^* = w_1\bar{u}_1 + w_2\bar{u}_{2m}, \quad u = x, y. \tag{2.2}$$

According to the general rule, we define $W_j = N_j/N$ and $w_j = n_j/n$, $j = 1$ or 2. Sample statistics obtained from a_{2m} are subscripted by "$2m$", (*e.g.*, $\bar{u}_{2m} = (1/m)\Sigma_{a_{2m}} u_i$); those from a_1 are subscripted by "1", (*e.g.*, $\bar{u}_1 = (1/n_1)\Sigma_{a_1} u_i$), and those for the first phase sample a' will be superscripted by a prime (*e.g.*, $\bar{x}' = (1/n')\Sigma_{a'} x_i$).

A large sample first order approximation to the variance of d^*, obtained by using the Taylor linearization, is given by

$$V(d^*) \cong \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_r^2 + \frac{W_2(k-1)}{n}S_{2r}^2 \tag{2.3}$$

where,

$$S_r^2 = S_y^2 + R^2S_x^2 - 2RS_{xy},$$

$$S_{2r}^2 = S_{2y}^2 + R^2S_{2x}^2 - 2RS_{2xy}, \tag{2.4}$$

R is the population ratio of $\bar{Y}$ to $\bar{X}$. S_u^2 and S_{2u}^2 are, respectively, the variance for the whole population and the population variance for the stratum of non-respondents of the variable u. S_{xy} and S_{2xy} are the covariances for the whole population and the population of non-respondents respectively.

The variance of d^* can be approximately estimated by

$$v(d^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_r^2 + \frac{w_2(k-1)}{n}\hat{S}_{2r}^2 \tag{2.5}$$

where,

$$\hat{S}_y^2 = \frac{1}{n-1}\left\{\sum_{a_1} y_i^2 + k\sum_{a_{2m}} y_i^2 - n\bar{y}^{*2} + w_2(k-1)s_{2my}^2\right\},$$

$$\hat{S}_r^2 = \frac{1}{n-1}\left\{\sum_{a_1}(y_i - r^*x_i)^2 + k\sum_{a_{2m}}(y_i - r^*x_i)^2\right\} \text{ and}$$

$$\hat{S}_{2r}^2 = \frac{1}{m-1}\sum_{a_{2m}}(y_i - r^*x_i)^2. \tag{2.6}$$

Note that $\hat{S}_y^2$ is an unbiased estimator of S_y^2. It seems natural to use $\hat{S}_r^2$ to estimate S_r^2 since the expression obtained from $\hat{S}_r^2$ by replacing r^* with R is a consistent estimator of S_r^2. The same argument can be used to justify the use of $\hat{S}_{2r}^2$.

An alternative estimator of $V(d^*)$ can be obtained by replacing $\hat{S}_r^2$ and $\hat{S}_{2r}^2$ with

$$\tilde{S}_r^2 = \hat{S}_y^2 + r^{*2}s_x'^2 - 2r^*s_{xy}^* \text{ and}$$

$$\tilde{S}_{2r}^2 = s_{2my}^2 + r^{*2}s_{2x}^2 - 2r^*s_{2mxy}, \tag{2.7}$$

respectively, in (2.5), where,

$$s_x'^2 = \frac{1}{n'-1}\sum_{a'}(x_i - \bar{x}')^2,$$

$$s_{2my}^2 = \frac{1}{m-1}\sum_{a_{2m}}(y_i - \bar{y}_{2m})^2,$$

$$s_{2x}^2 = \frac{1}{n_2-1}\sum_{a_2}(x_i - \bar{x}_2)^2,$$

$$s_{2mxy} = \frac{1}{m-1}\left(\sum_{a_{2m}} x_i y_i - m\bar{x}_{2m}\bar{y}_{2m}\right)$$

and s_{xy}^* is as in (2.9). This alternative estimator is likely to have a smaller variance than the estimator in (2.5) since the estimators $s_x'^2$ and s_{2x}^2 are based on larger samples and therefore more precise.

2.3 The Double Sampling Regression Estimator

We define the regression estimator by

$$t^* = \bar{y}^* + \hat{\beta}^*(\bar{x}' - \bar{x}^*) \tag{2.8}$$

where $\hat{\beta}^*$ is an estimator of $\beta = S_{xy}/S_x^2$. There could be several choices for $\hat{\beta}^*$, but a natural choice would be given by $\hat{\beta}^* = s_{xy}^*/s_x^{*2}$, where

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{a_1} x_i y_i + k\sum_{a_{2m}} x_i y_i - n\bar{x}\bar{y}^*\right) \text{and}$$

$$s_x^{*2} = \frac{1}{n-1}\left(\sum_{a_1} x_i^2 + k\sum_{a_{2m}} x_i^2 - n\bar{x}\bar{x}^*\right). \tag{2.9}$$

It is easy to show that s_{xy}^* and s_x^{*2} are unbiased for S_{xy} and S_x^2 respectively. An approximate variance of t^* is given as

$$V(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_l^2 + \frac{W_2(k-1)}{n}S_{2l}^2 \tag{2.10}$$

where S_l^2 and S_{2l}^2 are obtained from (2.4) by replacing R with β.

To estimate $V(t^*)$ we can use the following formula:

$$v(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_l^2 + \frac{w_2(k-1)}{n}\hat{S}_{2l}^2 \tag{2.11}$$

where,

$$\hat{S}_l^2 = \frac{1}{n-1}\left\{\sum_{a_1}(y_i - y_i^*)^2 + k\sum_{a_{2m}}(y_i - y_i^*)^2\right\},$$

$$\hat{S}_{2l}^2 = \frac{1}{m-1}\sum_{a_{2m}}(y_i - y_i^*)^2 \text{ and}$$

$$y_i^* = \bar{y}_i^* - \hat{\beta}^*(x_i - \bar{x}^*). \tag{2.12}$$

Like (2.7), a slightly improved estimator of $V(t^*)$ can be obtained by using

$$\tilde{S}_l^2 = \hat{S}_y^2 + \hat{\beta}^{*2}s_x'^2 - 2\hat{\beta}^* s_{xy}^* \text{ and}$$

$$\tilde{S}_{2l} = s_{2my}^2 + \hat{\beta}^{*2}s_{2x}^2 - 2\hat{\beta}^* s_{2mxy}. \tag{2.13}$$

3. CHOICE OF SAMPLING FRACTIONS

We shall now deduce the optimum k, n, and n' that minimize the variances of the proposed estimators for a specified cost, or that minimize the cost for a specified variance.

Let's consider a cost function for d^* given by

$$C = c'n' + cn + c_1 n_1 + c_2 m \tag{3.1}$$

where the c's are the costs per unit defined as follows:

- c' : the unit cost associated with the first phase sample, a';
- c : the unit cost of the first attempt on y with the second phase sample, a;
- c_1 : the unit cost for processing the respondent data on y at the first attempt in a_1;
- c_2 : the unit cost associated with the sub-sample, a_{2m} of a_2.

Since the value of n_1 is not known until the first attempt is made, the expected cost will be used in the minimization. The expected cost is given by

$$E(C) = C^* = c'n' + \left(c + c_1 W_1 + \frac{c_2 W_2}{k}\right)n. \tag{3.2}$$

The optimum values of k, n, and n' that minimize the variance of d^* for a fixed expected cost C^* are obtained by using Langrange multiplier. The optimum values thus obtained are:

$$k_o = \sqrt{\frac{c_2\left(S_r^2 - W_2 S_{2r}^2\right)}{S_{2r}^2(c + c_1 W_1)}}$$

$$n_o = \frac{C^*\sqrt{A}}{D\sqrt{G}} \text{ and } n_o' = \frac{C^*\sqrt{S_y^2 - S_r^2}}{D\sqrt{c'}} \tag{3.3}$$

where

$$A = S_r^2 + W_2(k_o - 1)S_{2r}^2,$$

$$G = c + c_1 W_1 + \frac{c_2 W_2}{k_o} \text{ and}$$

$$D = \sqrt{(S_y^2 - S_r^2)c'} + \sqrt{AG}.$$

If we let $\gamma = c_2/(c + c_1 W_1)$, $\delta = S_r^2/S_{2r}^2$ and $\xi = S_y^2/S_r^2$, then we have

$$k_o = \sqrt{\gamma(\delta - W_2)},$$

$$n_o = \frac{C^*\sqrt{1 + W_2(k_o - 1)/\delta}}{\sqrt{Gc'(\xi - 1)} + G\sqrt{1 + W_2(k_o - 1)/\delta}} \text{ and}$$

$$n_o' = \frac{C^*\sqrt{\xi - 1}}{c'\sqrt{\xi - 1} + \sqrt{Gc'\{1 + W_2(k_o - 1)/\delta\}}}. \tag{3.4}$$

The optimum values n_o and n_o' are proportional to the expected cost, C^*. To get the optimum values of k, n, and n' that, minimize $V(t^*)$ we simply substitute S_r^2 and S_{2r}^2 in the above expression in (3.3) with S_l^2 and S_{2l}^2, respectively. Table 1 shows optimum values of k_o, n_o, and n_o' for given parameters.

Table 1
Optimum Values of k_o, n_o, and n_o'

C^*	c'	c	c_1	c_2	δ	ξ	W_2	γ	k_o	G	n_o	n_o'
200	0.1	0.5	1	2	1	2	0.3	1.67	1.08	1.76	92	382
200	0.1	0.5	1	2	1	4	0.3	1.67	1.08	1.76	81	580
200	0.1	0.5	1	2	2	2	0.3	1.67	1.68	1.56	104	389
200	0.1	0.5	1	2	2	4	0.3	1.67	1.68	1.56	91	590
200	0.1	0.5	1	4	1	2	0.3	3.33	1.52	1.99	83	345
200	0.1	0.5	1	4	1	4	0.3	3.33	1.52	1.99	74	531
200	0.1	0.5	1	4	2	2	0.3	3.33	2.38	1.70	96	361
200	0.1	0.5	1	4	2	4	0.3	3.33	2.38	1.70	85	553
200	0.5	0.5	1	2	1	2	0.3	1.67	1.08	1.76	85	250
200	0.5	0.5	1	2	1	4	0.3	1.67	1.08	1.76	72	366
200	0.5	0.5	1	2	2	2	0.3	1.67	1.68	1.56	96	255
200	0.5	0.5	1	2	2	4	0.3	1.67	1.68	1.56	80	372
200	0.5	0.5	1	4	1	2	0.3	3.33	1.52	1.99	78	228
200	0.5	0.5	1	4	1	4	0.3	3.33	1.52	1.99	67	338
200	0.5	0.5	1	4	2	2	0.3	3.33	2.38	1.70	89	238
200	0.5	0.5	1	4	2	4	0.3	3.33	2.38	1.70	76	351

4. COMPARISON OF THE ESTIMATORS

In this section, the theoretical comparison of the performances of the proposed estimators with respect to the Hansen and Hurwitz (1946) estimator is made first without taking the cost into consideration and then with taking it into account.

4.1 Without Considering the Cost

The variance of the Hansen-Hurwitz estimator is

$$V(\bar{y}^*) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k-1)}{n}S_{2y}^2 \tag{4.1}$$

where $\bar{y}^*$ is defined as in (2.2).

$$V(\bar{y}^*) - V(d^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(2RS_{xy} - R^2S_x^2\right) + \frac{W_2(k-1)}{n}\left(2RS_{2xy} - R^2S_{2x}^2\right). \tag{4.2}$$

This is positive (*i.e.*, d^* is more efficient than $\bar{y}^*$) if $R < 2\beta$ and $R < 2\beta_2$, where $\beta_2 = S_{2xy}/S_{2x}^2$. On the other hand, we have

$$V(\bar{y}^*) - V(t^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{S_{xy}^2}{S_x^2} + \frac{W_2(k-1)}{n}\beta S_{2x}^2(2\beta_2 - \beta). \tag{4.3}$$

Therefore, t^* is more efficient than the Hansen-Hurwitz estimator if (4.3) is positive. One particular condition under which this can occur is that $\beta_2 \geq \beta/2$ with $\beta \geq 0$. The conditions we discuss here are sufficient and thus, d^* or t^* can be more efficient than $\bar{y}^*$ under more relaxed conditions.

4.2 Considering the Cost

We shall now compare the proposed estimators with the Hansen-Hurwitz estimator ($\bar{y}^*$) making use of the cost function given in section 3.

For the estimator $\bar{y}^*$, if a straight random sample is taken (without using double sampling procedure) for y, the optimum sample size for an expected cost,

$$C^* = \left(c + c_1 W_1 + \frac{c_2 W_2}{k}\right)n,$$

similar to the one in (3.2) can be obtained by the same technique (*i.e.*, Langrange multiplier) used in section 3 as follows:

$$n_{oHH} = \frac{C^*}{c + c_1 W_1 + c_2 W_2 / k_{oHH}} \text{ and}$$

$$k_{oHH} = \sqrt{\frac{c_2(S_y^2 - W_2 S_{2y}^2)}{S_{2y}^2(c + c_1 W_1)}}. \quad (4.4)$$

Then the optimum variance of the Hansen-Hurwitz estimator becomes

$$V(\bar{y}^*) = \left(\frac{1}{n_{oHH}} - \frac{1}{N}\right) S_y^2 + \frac{W_2(k_{oHH} - 1)}{n_{oHH}} S_{2y}^2. \quad (4.5)$$

If we compare this with $V(d^*)$ with the optimum choices of k, n, and n' in (3.3), then the condition that d^* will be more precise than $\bar{y}^*$ is given by

$$2\rho - Rh > \frac{1}{1 - \theta_1} \times \left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2 Q_{HHy}) - hQ_x(2\beta_2 - R)\right\} \quad (4.6)$$

where

$$h = \frac{S_x}{S_y}, \theta_1 = \frac{n_o}{n_o'}, \theta_2 = \frac{n_o}{n_{oHH}}, Q_{HHy} = \frac{W_2(k_{oHH} - 1)S_{2y}^2}{S_y^2},$$

$$Q_u = \frac{W_2(k_o - 1)S_{2u}^2}{S_u^2}, u = x, y,$$

and ρ is the correlation coefficient between x and y.

We can obtain a similar comparison between $\bar{y}^*$ and t^*. That is, t^* is more efficient than $\bar{y}^*$ if

$$2\rho - \beta h > \frac{1}{1 - \theta_1} \times \left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2 Q_{HHy}) - hQ_x(2\beta_2 - \beta)\right\}. \quad (4.7)$$

4.3 Empirical Comparison of the Proposed Estimators

The relative efficiencies of the estimators d^* and t^* with respect to $\bar{y}^*$ are compared using an artificially generated population. The parameters of the population are:

$R = 1.92, \beta = 1.52, \rho = 0.85, R_2 = 1.88, \beta_2 = 1.47,$

$\rho_2 = 0.83, N = 1{,}000, N_2 = 302, S_x^2 = 766.54,$

$S_y^2 = 2426.82, S_{xy} = 1164.08, S_{2x}^2 = 433.63,$

$S_{2y}^2 = 1350.05$, and $S_{2xy} = 638.32$.

The relative efficiencies of d^* and t^* are presented in Table 2. Note that R is substantially different from β, which means that the regression line does not pass through the origin. Under this population, the regression estimator t^* is more efficient than the ratio estimator d^*. We notice also that the optimum initial sample size, n_o' for t^* is more than for the estimator d^*. The reverse is the case for the optimum second phase sample size n_o. This is so because the regression estimator can be more precise with a smaller second phase sample size so that it allows to allocate more to the first phase sample. Finally the optimum inverse sampling rate k_o is practically the same for the two estimators.

When the linear regression line passes through the origin, the advantage of t^* over d^* disappears, as expected and confirmed in another empirical comparison not shown here.

Table 2
The Relative Efficiencies of d^* and t^* with Respect to $\bar{y}^*$ ($C^* = 200$, $c = 0.5$, $c_1 = 1$)

c'	c_2	k_{oHH}	n_{oHH}	k_o	n_o	n_o'	Efficiency
				d^*			
0.1	2	1.58	127	1.46	92	514	1.85
0.1	4	2.23	115	2.06	85	477	1.91
0.3	2	1.58	127	1.46	78	250	1.23
0.3	4	2.23	115	2.06	73	234	1.32
				t^*			
0.1	2	1.58	127	1.47	89	563	2.11
0.1	4	2.23	115	2.08	83	523	2.19
0.3	2	1.58	127	1.47	74	269	1.34
0.3	4	2.23	115	2.08	70	253	1.45

5. CONCLUSIONS

We proposed ratio and regression estimators based on the double sampling procedure when there is non-response on the main character and the population mean of the auxiliary variable is not known. The potentially serious non-response bias is eliminated by sub-sampling the non-respondents as in the Hansen and Hurwitz procedure (1946). We derived optimum sample sizes for a given set of unit costs and compared theoretically and empirically the performance of our estimators with that of the Hansen and Hurwitz estimator.

When there is a strong linear relationship between the main character and the auxiliary character and the auxiliary data can be collected cheaply with a large sample size, our estimators are substantially superior to the Hansen and Hurwitz estimator. Our procedure can be useful when there is a serious concern about the nonresponse bias that is difficult to handle with the usual weighting adjustment or imputation.

ACKNOWLEDGEMENT

We are grateful to the referees and associate editors for their comments that helped to improve our paper.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.

HANSEN, M. H., and HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

OH, H.L., and SCHEUREN, F.J. (1983). Chapter 13. Weighting adjustment for unit non-response. In *Incomplete Data in Sample Surveys*, (I. Olkin, W.G. Madow, and D.B. Rubin, Eds.). Theory and Bibliographies, 2, 143-184. New York: Academic Press.

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 189-198
Statistics Canada

Modeling Interviewer Effects in Panel Surveys: An Application

JAN PICKERY and GEERT LOOSVELDT[1]

ABSTRACT

In this paper we will combine two applications of multilevel models. The multilevel model is suitable to analyze interviewer effects on survey data. It can also be used to analyze longitudinal – "repeated measurements" – data. We will analyze a data quality indicator of panel data that come from the Belgian Election Studies. These panel data consist of only two waves. The respondents that cooperated twice are for the most part not interviewed by the same interviewers. This results in a complex data structure with measurements nested in respondents, and respondents nested in interviewers, but without an overall hierarchical nesting structure: cross-classification. This complicated data structure will be analyzed in two different ways: an analysis of all respondents and an analysis of only those who are interviewed twice by the same interviewer. The results of these different analyses will be compared. We conclude that the multilevel cross-classified model is a very flexible and useful tool to analyze interviewer effects in panel surveys.

KEY WORDS: Multilevel models; Cross-classifications; Panel surveys; Interviewer effects; Don't know answer.

1. INTRODUCTION

In this paper we analyze the effect of respondent and interviewer characteristics on the number of "don't know" answers in two waves of the panel survey from the Belgian Election Studies. We use different multilevel models for a subset of the dataset and for the entire dataset. The main purpose of the article is to illustrate how interviewer effects in a panel survey can be analyzed using multilevel models.

A multilevel or hierarchical model is an appropriate tool to analyze data with nested structures, *e.g.*, pupils nested in schools or patients in hospitals. A multilevel model can include variables of the different levels of nesting, but it also takes account of the variability associated with each level. The typical quality of the models is not the functional form relating the variables of the different levels, but rather a more sophisticated treatment of the error structure (DiPrete and Forristal 1994, 334). In education research for instance a multilevel model can account for variation between schools and variation between pupils. Moreover the model tries to replace this variance attributed to both levels by variables of either level. These models are described in various textbooks like Bryk and Raudenbush (1992), Goldstein (1995), Kreft and de Leeuw (1998) and Snijders and Bosker (1999).

Multilevel or hierarchical models also offer the best possibilities to analyze interviewer effects on survey data (Hox 1994). A hierarchical model is the best tool to tackle the "respondents nested within interviewers" – design. Other statistical techniques require mutual independence of interviewer and respondent characteristics, which is – most of the time – not the case because of the hierarchical structure of the data. In a multilevel model both the regression coefficients and the variance components are conditional on the explanatory variables in the model, which is a useful property if there is no complete orthogonalization of interviewer and respondent variables (Hox 1994, 307). When respondents are not randomly assigned to interviewers, respondent and interviewer characteristics can become confounded since respondents from a specific area will most likely be interviewed by interviewers from the same area. In such a situation, if the relevant respondent variables are put in the multilevel model, interviewers are equalized by statistical means. For that reason the assumptions of an analysis of interviewer effects with a multilevel model are more realistic than those of an ANOVA or ANCOVA. Furthermore the hierarchical model allows estimation of both the interviewer variance and the effects of explanatory variables measured at the interviewer and the respondent model. This possibility of replacing variance attributed to respondents/interviewers with the effects of respondent/interviewer characteristics allows for wider generalizations.

The multilevel model can also fruitfully be used to analyze longitudinal – "repeated measurements" – data (see *e.g.*, Goldstein 1995, 87-95; Snijders 1996 and Yang and Goldstein 1996). There are alternatives to analyze the "measurements nested in individuals" – design, but multilevel analysis has some clear advantages. Using a hierarchical model, it is feasible to handle unbalanced designs – not all individuals have the same number of measurements – and quite easy to incorporate changing covariates. Besides, the model allows more nesting levels. The individuals can be nested in another higher level unit.

We will analyze respondent and interviewer effects on the number of "don't know" answers on a series of questions regarding political parties in a panel survey. We have

[1] Jan Pickery and Geert Loosveldt, Department of Sociology, University of Leuven, E. Van Evenstraat 2B, 3000 Leuven, Belgium. E-mail: jan.pickery@soc.kuleuven.ac.be; geert.loosveldt@soc.kuleuven.ac.be.

measurements (wave 1 and wave 2) nested in respondents (the longitudinal design) and respondents nested in interviewers. Our panel data consist of two waves. During the second wave the respondents are for the most part not interviewed by the same interviewers. The purely hierarchical nesting has broken down and a more complex data structure is the result. To handle this data structure it is necessary to conceive the measurements as being nested into two different classification structures: measurements nested in respondents and in interviewers. This is called a cross-classified design, because the nesting of the levels is not purely hierarchical.

In this article we'll start with a simple data structure and the appropriate model. Afterwards the model will become more complex. We'll perform two analyses. In our first analysis we work only with the respondents who are interviewed twice by the same interviewers. Afterwards we will analyze all the respondents, including those who were interviewed only once. In the first analysis the purely hierarchical structure remains intact. The model is a "simple" three level one: measurements nested in respondents nested in interviewers. Analysis 2 sets up a cross-classified model. In that model the measurements are classified by respondents and interviewers.

The next section reflects on the nature of our dependent variable, the "don't know" answer, and the way to analyze it. In the third section we'll describe our data in detail to clarify the complex structure. The following section (4) treats in brief the different models that we will combine. Section 5 presents the variables in our analysis. In sections 6 and 7 we discuss the setup of our 2 different models and report the results of the analyses. Section 8 concludes the article.

2. ANSWERING "DON'T KNOW"

It has become generally acknowledged that the use of a "don't know" or a "no opinion" filter increases the proportion of respondents who give this answer, and that the increase itself is a function of the nature of the filter used (Schuman and Presser 1981, 143). Krosnick argues that answering "don't know" is one form of satisficing. Satisficing occurs when a respondent is not motivated to expend the mental effort necessary to generate optimal answers. A "no opinion" answer is an acceptable answer but it is the result of a "weak" cognitive process. Satisficing is a function of task difficulty, and the respondent's lack of knowledge, ability and motivation. This theoretical reasoning is consistent with the finding that offering a "don't know" response option increases the proportion of respondents who select it, particularly among respondents with little formal education and people who consider an issue to be less personally important. (Krosnick 1991). Following this argumentation, answering "don't know" is mainly explained by respondent characteristics that can be related to the cognitive aspect of answering questions. Previous research points us to the following characteristics of interest: education (*e.g.*, Sudman and Bradburn 1974), age (see *e.g.*, Groves 1989, 441-443), sex (*e.g.*, Hox, de Leeuw and Kreft 1991), and a measure of involvement or interest in the subject (*e.g.*, Groves 1989, 419).

However answering questions is not only a cognitive process of the respondent but it is also a communicative process (Schwarz and Sudman 1995). Within this process the interviewer plays an important role. There is a lot of literature about the interviewer as a source of survey measurement error (Groves 1989). The main idea is that interviewers are not "neutral" collectors of data but that they can influence the respondents' answers. Item non-response too is subject to interviewer effects as has been shown long ago by *e.g.*, Hanson and Marks (1958) and Bailar, Bailey and Stevens (1977). A social scientist interested in the explanation of "don't know" answers should therefore include respondents and interviewers in the analysis. The number of "don't know" answers will be the dependent variable of our analyses.

3. DESCRIPTION OF THE DATA STRUCTURE

After the 1991 General Election in Belgium a national survey was set up in which 4,544 face to face interviews were conducted in the three Regions in the early months of 1992. A two-stage self-weighting sample (see *e.g.*, Särndal, Swensson and Wretman 1992, 141-144) was used. The sample was representative for the population of 18-74 years old (ISPO/PIOP 1995). In this article we will use the data from the Flemish region, which cover 2,691 Flemish respondents, interviewed by 163 interviewers (Carton, Swyngedouw, Billiet and Beerten 1993). After the 1995 Elections a similar survey was set up. Due to budgetary constraints the sample had to be smaller for the second wave. So the 2,691 respondents were used as a group to sample from and, in second order, there had to be new respondents to compensate for the aging of the youngest cohort from 1991. Finally 2,099 respondents were interviewed by 167 interviewers. This sample contained 1,762 panel respondents and 337 new respondents (see Beerten, Billiet, Carton and Swyngedouw 1997 for a detailed technical report of the sample plan). Only 55 of the interviewers of the first wave collaborated again. So there were 112 new interviewers in the second wave.

This gives us a dataset with 3028 respondents (2,691 + 337) and 275 interviewers (163 + 112). For 1,762 respondents we have a measurement in both waves, for the rest (1266) there is only one measurement. The structure of the dataset can be represented in a table similar to table 1 (see also Goldstein 1995, 114). Each x in the table reflects an observation. The complete dataset contains 4,790 observations ((1,762 × 2) + 1,266). Each type of respondent in the table represents a possible occurrence in the dataset.

Table 1
A Representation of the Dataset

	Respondent Type 1		Respondent Type 2		Respondent Type 3	Respondent Type 4	N (Interviewers)
Wave	1	2	1	2	1	2	
Interviewer Type A	x	x	x			x	47
Interviewer Type B				x	x		8
Interviewer Type C			x		x		108
Interviewer Type D				x		x	112
N (Respondents)	374		1388		929	337	

This table illustrates that we have three kinds of respondents: panel respondents who are interviewed twice by the same interviewer (Type 1), respondents who cooperated twice but were interviewed by different interviewers (Type 2) and respondents who are only interviewed once (Type 3 and 4). Our 2 different analyses are based on these different types of respondents. In Analysis 1 we'll look at the respondents whose situation corresponds with that of Type 1. Only 374 Respondents satisfy this condition. They were interviewed twice by the same interviewer. Analysis 2 takes all the 3028 respondents into account (Respondents 1 to 4 of the table).

Furthermore the table shows that we can also discern different interviewers: interviewers who collaborated twice (Type A and B) and interviewers who collaborated only the first wave (Type C) or only the second wave (Type D). The interviewers of Type B collaborated in both waves, but never interviewed the same respondents twice (unlike the interviewers of Type A).

To analyze this complex data structure we will combine three different models that are presented in the next section.

4. A SHORT DESCRIPTION OF THE DIFFERENT MULTILEVEL MODELS USED IN THE ANALYSES

4.1 The General Multilevel Model

The first model we need is the general multilevel model, which has the following form:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}, \quad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \text{ and } \beta_{1j} = \beta_1 + u_{1j} \quad (2)$$

or

$$\beta_{0j} = \beta_0 + \gamma_{01} z_{1j} + u_{0j} \text{ and } \beta_{1j} = \beta_1 + \gamma_{11} z_{1j} + u_{1j}. \quad (3)$$

Subscript i refers to the level 1 unit and subscript j to the level unit 2. In our situation level 1 indicates the respondent and level 2 the interviewer. So the response variable Y of respondent i, interviewed by interviewer j is dependent on the x variable of that respondent. This relationship looks like an ordinary regression model but the parameters of the model are interviewer specific. The β's differ from interviewer to interviewer. For each β, there is an interviewer residual (u_{0j} or u_{1j}). The β's can also be made dependent on higher level variables (interviewer characteristics), allowing for generalization across interviewers. We have one second level variable z. Substituting (3) into (1) results in the following overall model:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \gamma_{01} z_{1j} + \gamma_{11} z_{1j} x_{1ij} + u_{1j} x_{1ij} + u_{0j} + e_{ij}. \quad (4)$$

Of course more x and z variables can be included in these relationships. We assume that the residuals u_{0j}, u_{1j} and e_{ij} have means 0 given the values of the explanatory variables z and x. Furthermore it is assumed that the level 1 residuals (e_{ij}) are independent. The level 2 residuals (u_{0j} and u_{1j}) are assumed to be independent from e_{ij} and to have a joint multivariate normal distribution with covariance matrix Ω. They don't have to be independent from each other. Usually they are correlated.

4.2 The Multilevel Model for Longitudinal Analysis

The second model we need is the longitudinal multilevel model. In an analysis of a "repeated measurements" – design with a hierarchical model, the measurements are considered to be the first level and the individual the second. Most of the time the individual units will be persons, but of course they can be other units, like *e.g.*, schools or countries. In our analysis the individuals are the respondents. The analysis tries to estimate a growth curve on the base of the different measurements and to compare differences in curves given individual characteristics. Each observed value is made conditional upon the time of measurement – which can be a measure of time, but also age – and possible transformations of this measurement. Usually the curve is assumed to be a polynomial, which has the following form:

$$Y_{it} = \pi_{0i} + \pi_{1i} t + \pi_{2i} t^2 + \ldots + \pi_{ki} t^k + e_{it}. \quad (5)$$

Y_{it} is the observed value for respondent i on moment t, t can be time of measurement or age. π_{hi} ($h = 0 \ldots k$) are the trajectory parameters or growth parameters for subject i, k is the degree of the polynomial. In a simple case k has the value 1 and then there is a linear curve. If there are m moments of measurement, a polynomial with degree of $m - 1$ will result in an exact reproduction of the curve. Of course it is more interesting to use a polynomial with a lower degree if that yields a satisfactory reproduction of the curve. You can test whether the model with degree $k + 1$ results in a significant improvement compared to the model with degree k.

The growth parameters have also a subscript for the individual (respondent). The model states that these parameters differ from individual to individual. The second part of the model defines these parameters:

$$\pi_{0i} = \pi_0 + r_{0i} \quad (6)$$

or

$$\pi_{0i} = \pi_0 + \beta_{01} x_{1i} + r_{0i}. \quad (7)$$

The individual parameter equals a general parameter (π_0) + an individual residual (r_{0i}). By the inclusion of individual characteristics (x) it may be possible to reduce the individual specific part, thus generalizing across respondents. In line with (1) and (2) we chose x to denote the individual (respondent) characteristics. But it is worth mentioning that in this model the x variables are higher level (level 2) variables. The individual characteristics can be fixed (the same for all moments of measurement) or varying.

4.3 Cross-Classified Models

The third model we will use is the cross-classified model. Not all data structures are purely hierarchical. Units may be classified along more than one dimension (see Goldstein 1995, 113-116). For example students can be classified by the school they go to and by the neighborhood they live in. In our example measurements are classified by respondents and by interviewers. A cross-classified model has the following form (subscripts j_1 and j_2 refer to the 2 different classification structures):

$$Y_{ij_1j_2} = \beta_{0j_1j_2} + \beta_{1j_1j_2} x_{1ij_1j_2} + e_{ij_1j_2}, \quad (8)$$

$$\beta_{0j_1j_2} = \beta_0 + u_{0j_1} + u_{0j_2} \text{ and } \beta_{1j_1j_2} = \beta_{1j_1j_2} + u_{1j_1} + u_{1j_2}. \quad (9)$$

Equation (9) can be reformulated the same way as equation (3).

$Y_{ij_1j_2}$ is the observed value for individual i, classified by j_1 and j_2. In our case: the observed value for measurement i on respondent j_1, interviewed by interviewer j_2. The parameters associated with the independent variable x have a residual for both classifying structures. For this model the additional assumption is made that the residuals of the different classifying structures (in our case: the respondent and interviewer residuals) are mutually independent (u_{0j_1} and u_{1j_1} versus u_{0j_2} and u_{1j_2}).

Raudenbush (1993) discusses this kind of models and the use of the EM algorithm to estimate them. Rasbash and Goldstein (1994) and Goldstein (1995, 123-124) show how these models can be specified and estimated using a purely hierarchical formulation and (consequently) standard multilevel software. The way to do this is to specify one of the classifications as a standard hierarchical one, then define a dummy for each unit of the other classification, specify that each of these dummy variables has a random coefficient at the higher level and constrain the resulting sets of variances to be equal.

In section 6 and 7 we'll use these 3 different models. They can all be implemented in MLn/MLwiN, software for multilevel modeling. Firstly we take a closer look at the variables we will use in the analysis.

5. VARIABLES IN OUR ANALYSIS

One of the more difficult tasks during the interview of the election study was rating six parties on different 11-point scales. Three scales were presented to the respondents: catholicism, economic liberalism and federalism. An explicit "don't know" filter was included in the question, but it was not mentioned on the card with the alternatives given to the respondent. The entire question is included in the Appendix. We expected a considerable number of "don't know" answers because of the degree of complexity of the task. The explicit filter was expected to raise that number as well (see *e.g.*, Schuman and Presser 1981).

In the first wave the average number turned out to run up to more than 4 "don't know" answers per respondent. Almost 20% of the respondents made use of this possibility at least 9 times out of the 18. If we consider only the panel respondents the mean number is a bit lower (3.8). This is not surprising since we could expect that "multi-users" of the "don't know" answer would be underrepresented in the second wave because of lack of interest in the subject of the survey and/or difficulties in answering the questions. In the second wave the overall mean is 3.6 and the mean for the panel respondents 3.4. The numbers for the respondents that were interviewed twice by the same interviewer are 3.9 and 4.2 respectively. There is no explanation why the number of "don't know" answers during the second wave is higher than the overall mean for these respondents.

At the measurement level we'll use the year of the interview as indication of time of measurement. We've recoded this variable, so time has the value 0 for the first wave and 3 for the second wave.

At the respondent level we have 3 independent variables: sex (0 = man, 1 = woman), completed education (0 = low, 1 = high) and the extent to which the respondents follow political news in the press (press: 1 = (almost) always - 5 never). The first 2 variables are constant for the 2 times of measurement. The third is a time-varying covariate and the question phrasing also slightly changed for the second survey. The two different questions are also included in the appendix. The dissimilarity in the phrasing induces an additional difficulty in setting up the model. The way to handle such a variable is to standardize it (mean 0, variance 1) for each time of measurement and (afterwards) to ascribe the value 0 to the time of measurement when the question wasn't asked. The reference value for those variables is their mean (see Snijders 1996, 422). This gives us 2 variables: press1 for the first occasion and press2 for the second. The former has the value 0 for all respondents for the second measurement and the latter for the first measurement. We don't take up the respondent's age in the model,

since this variable would correlate too much with the time measurement at the occasion level.

In order not to complicate the analysis too much we don't take up interviewer variables. We just assume there is an interviewer effect, without trying to explain that effect in terms of interviewer characteristics.

6. ANALYSIS 1: RESPONDENT TYPE 1 OF TABLE 1

The first analysis considers only those respondents who are interviewed twice by the same interviewer (cfr. Type 1 from Table 1). This analysis requires a "simple" three level model: measurements nested in respondents nested in interviewers. The hierarchical structure is unambiguous. This model is similar to the example in chapter 8 in the Bryk and Raudenbush book (1992). In that example the authors analyze the progress in academic achievement of students in schools.

Our dependent variable is the number of "don't knows"s for respondent i on moment t, interviewed by j (Y_{tij}). We have only 2 measurements so the degree of the polynomial cannot exceed 1. This results in the following level 1 equation:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\,\text{YEAR} + e_{tij}.$$

Our time variable (t) is the year of the interview which has the value 0 (1992) or 3 (1995). We will test whether π_{1ij} is significant. If not, this leaves us a null model or "naive" model (see Snijders 1996, 411), in which the number of "don't know" answers doesn't change over time, and both measurements can be considered as retests of the same constant value. The coefficients in the level 1 equation are respondent and interviewer specific.

At the respondent level we'll include 3 variables: sex, education and the 2 press variables. So our level 2 equation contains 4 variables:

$$\pi_{0ij} = \pi_{0j} + \beta_{01j}\,\text{SEX}_i + \beta_{02j}\,\text{EDUCATION}_i + \beta_{03j}\,\text{PRESS1}_i + \beta_{04j}\,\text{PRESS2}_i + r_{0ij}.$$

If the parameter estimate associated with year is significant we'll have a similar equation for π_{1ij}.

At the third level (interviewer) we won't include anymore variables, but we will fit a random intercept and random slopes. So we have the following level 3 equations:

$$\pi_{0j} = \pi_0 + u_{0j} \text{ and } \beta_{01j} = \beta_{01} + u_{01j}, \ldots$$

Implementing these model specifications in MLn gives us the following results.

Model a in the table is the null model. This is a model without independent variables, neither at the measurement level, nor at the respondent level. In this model there is no evolution in the number of "don't know" answers. But the variance of the dependent variable is divided in a measurement part, a respondent part and an interviewer part. All variances are significant. This indicates that there is between wave variation, that some respondents use the answer more than others and that some interviewers will get more "don't know" answers than other interviewers.

Table 2
Analysis of Respondents who were Interviewed Twice by the Same Interviewers (s.e. in brackets)

	model a	model b	model c	model d
Fixed				
Measurement level				
constant	4.136 (0.322)	4.028 (0.358)	3.749 (0.442)	3.754 (0.523)
year		0.072 (0.089)		
Respondent level				
sex			2.393 (0.434)	2.458 (0.414)
education			-1.675 (0.425)	-1.778 (0.446)
press1			0.911 (0.263)	0.887 (0.233)
press2			1.483 (0.236)	1.426 (0.234)
Random				
Interviewer level				
$\sigma^2_{\text{constant}}$	2.249 (1.040)	2.251 (1.043)	2.666 (0.969)	6.090 (2.109)
$\sigma_{\text{education/constant}}$				-4.099 (1.816)
$\sigma^2_{\text{education}}$				1.396 (1.819)
Respondent level				
$\sigma^2_{\text{constant}}$	14.470 (1.714)	14.480 (1.714)	8.939 (1.308)	8.692 (1.332)
Measurement level				
σ^2_e	13.320 (0.974)	13.300 (0.974)	13.270 (0.969)	13.250 (0.969)
-2 LL	4519.35	4518.62	4414.52	4395.68
Δ df*		1	4	6

Note: * compared to model a

The inclusion of the variable YEAR does not provide a better fit of the model. The decrease in deviance (-2 Log L) is not significant, neither is the parameter of the variable significant (model b). We can conclude that there is no significant overall evolution in the number of "don't know"s. We can go on with a model without the time variable.

The respondent variables do result in a considerable improvement of fit of the model. The decrease of the - 2 Log L value is large and clearly significant ($p<0.001$). According to the analysis women use the "don't know" alternative more than men do and highly educated respondents less than respondents with lower education. Following the political news in the press reduces your chance to answer "don't know". Both press1 and press2 are significant (model c). The inclusion of respondent variables also results in a substantial decrease of the variance at the respondent level.

We also tried to fit random slopes at the interviewer level (model d). Our analysis showed some variation in the parameter associated with the respondent's education. This is the only independent variable with a varying coefficient at the third level. $\sigma^2_{education}$ is not significant, but there is an important covariance between the residual for the constant and the residual for education ($\sigma_{education/constant} = -4.099$). The covariance is negative, indicating that interviewers with a higher constant have a smaller coefficient for education. Since the fixed parameter for education is negative, it will be even more negative for those interviewers, thus having a larger absolute value. Hence for interviewers who stimulate more "don't know" answers, the difference between less educated respondents and more educated respondents will be larger. In model d the value of $\sigma^2_{constant}$ at the interviewer level has increased considerably, compared to model c. In this model the variance at the interviewer level is dependent on the values of the explanatory variable education and it will be larger for zero values of education. That is another interpretation of model d: the variance between interviewers is much higher for lower educated respondents than for higher educated respondents. This model with a more complex variance structure at level 3 has a better fit than the previous models.

When including YEAR in model c or model d, it turned out to be not significant either. Also in our final models there is no evidence for an evolution in the number of "don't know"s between the two waves. All models prove a significant interviewer effect. But the relative size of the variance shows that there is more variation between respondents than between interviewers.

7. ANALYSIS 2: ALL RESPONDENTS

In this analysis we look at all the respondents: the panel respondents that were interviewed twice by the same interviewer, the other panel respondents and those who were interviewed only once. This second analysis breaks down the hierarchical structure. Measurements are still nested in respondents and respondents are still nested in interviewers. But there is no overall hierarchical structure, since the interviewer can (and most of the time will) change between the two waves (see section 3). Our dependent variable is still the number of "don't know"s of respondent i interviewed by j on moment t (Y_{tij}). But the model has changed. The level 1 equation hasn't:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\,\mathrm{YEAR} + e_{tij}.$$

In this notation we use π, since the level 1 model is also a growth curve. But this equation matches the level 1 model of the cross-classified model (equation (8), section 4.3). Furthermore we still use i for the respondent and j for the interviewer. But it is important to notice that this is not the same model as the one of analysis 1. These subscripts correspond to the j_1 and j_2 of equations (8) and (9).

There is no "real" third level. To fit the cross-classified model in MLn, we have to define a third level, but conceptually the respondent and interviewer are at the same level in this model. This leads to the following level 2 equation:

$$\pi_{0ij} = \pi_0 + \beta_{01}\,\mathrm{SEX}_i + \beta_{02}\,\mathrm{EDUCATION}_i + \beta_{03}\,\mathrm{PRESS1}_i + \beta_{04}\,\mathrm{PRESS2}_i + r_{0i} + r_{0j}.$$

The interviewer specific part (r_{0j}) is included in the second level, so there is no interaction between the interviewer variance and the respondent variables. This is the main difference with analysis 1.

A cross-classified model requires enormous computations. We have 3,026 respondents and 275 interviewers. This would mean 275 dummies with all varying coefficients at the artificial third level. Up till now it is impossible to fit such a model. The storage required by the worksheet is far too large (see Goldstein 1995, 118 and Rasbash and Woodhouse 1996, 85-86 for details). It is possible to reduce these storage requirements and improve the speed of model estimation by dividing the dataset in subsets in which the cross-classification implies fewer cells. In our case we look for separate groups of measurements that are classified by fewer respondents and interviewers. The analysis of 1 group of 1,000 measurements classified by 500 respondents and 100 interviewers is computationally more demanding than the analysis of a dataset consisting of 10 groups of 100 measurements each, classified by 50 respondents and 10 interviewers. Sometimes it is worth omitting some of the observations (measurements in combinations of respondents and interviewers that hardly occur) to make the partitioning more efficient.

MLn/MLwiN provides some procedures (via the commands XSEArch and BXSEarch) that are designed for that partition (Rasbash and Woodhouse 1996, 89-93). We used the BXSEarch command. The command starts an enhanced procedure, which attempts to provide the maximum separation with the minimum deletion of data. We started with

4,790 measurements, 3,026 respondents and 275 interviewers. After omitting the observations indicated by the BXSEarch command, we're left with 4,597 measurements on 3,026 respondents interviewed by 275 interviewers. No higher level units (respondents nor interviewers) are left out. The procedure resulted in 7 partitions with a maximum of 44 cells in the cross-classification of respondents and interviewers. The model converged sufficiently fast when implied this way.

The results of the analysis are reported in Table 3.

Table 3
Analysis of all the Respondents (s.e. in brackets)

	model a	model b	model c
Fixed			
Measurement level			
constant	3.894 (0.136)	3.967 (0.155)	3.864 (0.165)
year		-0.053 (0.055)	
Respondent level			
sex			1.808 (0.153)
education			-1.914 (0.148)
press1			1.185 (0.090)
press2			1.197 (0.102)
Random			
Level 2			
Interviewer			
$\sigma^2_{constant}$	2.777 (0.373)	2.716 (0.368)	2.844 (0.363)
Respondent			
$\sigma^2_{constant}$	11.810 (0.635)	11.800 (0.635)	7.017 (0.527)
Measurement level			
σ^2_e	13.130 (0.475)	13.150 (0.476)	13.460 (0.480)
-2 LL	27717.1	27716.3	27042.1
Δ df*			

Note: * compared to model a

This table looks very much the same as Table 2 but there is an important difference. In the random part we marked level 2 – interviewer and respondent to make clear that the interviewers do not constitute a third level in this analysis.

Model a is the null model: no explanatory variables, but the variance of the dependent variable separated in a measurement part, a respondent part and an interviewer part. There is a significant interviewer variance. Thus in this design we again have evidence for an interviewer effect. You have to be careful about the interpretation of the relative sizes of the variances if one classification has far fewer units than the other (Goldstein 1995, 117-118). It's not fully correct to state that there's 5 times as much variation between respondents than between interviewers, but again there is much more variability between respondents than between interviewers.

In the next model (model b) we've included the time variable (YEAR). Again this variable turns out to be not significant and its inclusion does not provide a better fit of the model. Again we can conclude that there is no significant overall evolution in "don't know" answers over time.

Model c is the model with the respondent variables. They are all significant and this model has a far better fit than the previous ones. The substantive interpretation of the parameters is the same as in analysis 1. Women use the "don't know" answer more than men and a higher education results in less "don't know"s. The extent to which the respondents follow the political news in the press is also a predictor of the use of the "don't know" answer. The less they follow politics the more they answer "don't know".

8. CONCLUSION AND DISCUSSION

The general conclusions of this article are methodological as well as substantive.

Our analysis confirms previous research findings about the use of the "don't know" answer. It is related to the respondent's education, sex and a measure of involvement or interest in the subject. Furthermore it is likely to diverge from interviewer to interviewer. All our analyses showed a significant interviewer effect. We did not find a significant evolution in the use of the "don't know" answer over time in the two waves of the survey. The interviewer effects prove that the "don't know" response alternative is not merely a result of the respondent answering the questions. It stresses the necessity of an interviewer training, which includes instructions on how to ask difficult questions and how to deal with "don't know" answers.

As in most panel surveys, the nonresponse in the second wave of this panel survey was not totally random. It is related to the respondent's living arrangement, his or her political interest and a few socio-demographic variables (Loosveldt and Carton 1997). This selective dropout puts limits to the generalizability of the results concerning the evolution in the dependent variable, but our analyses did not show a general evolution in the use of the "don't know" answer anyway. An impact of selective nonresponse in the second wave on the size of the interviewer effect is not unlikely either as interactions between the respondent characteristics and the interviewer effects are possible, as analysis 1 showed. But it is unlikely that this will affect the substantive conclusions about the interviewer effects. Given the results of analysis 1 and the conclusions in the Loosveldt and Carton paper, one could even expect that the interviewer effect in analysis 2 and consequently the overall interviewer effect might be somewhat underestimated. Loosveldt and Carton (1997, 1021) show that lower educated respondents are more likely to drop out of the survey than higher educated respondents and analysis 1 showed that the interviewer variance is higher for lower educated respondents.

The methodological conclusions consider the use of the different models to analyze interviewer effects in panel surveys. The analyses presented in this paper show that quite complex designs with complicated data structures can be analyzed by specifying the appropriate multilevel model.

The first model (Analysis 1) only suits in a tiny number of cases. It is not so common to ascribe the same interviewers to the same respondents for different waves of a panel survey, neither is it always feasible.

The second model (Analysis 2) is an appropriate tool but can require enormous computations. MLn is quite powerful and helps to decrease the storage requirements, at the cost of a small loss of information. Besides, the second model has its limitations too. Using this method it is not possible to model interactions between respondent variables and interviewer variance, as we did in the first analysis, or between respondent and interviewer variables. However the analysis showed that this model could be a very useful and flexible tool. The cross-classified model is also suitable when the number of measurements increases. A panel survey with 3 or 4 or even more waves, where some interviewers are retained and some are new at each occasion would require exactly the same analysis. The multilevel model also knows how to handle respondents for whom 1 or more measurements are missing, as our analysis showed. The pliability of this model outweighs the impossibility to include respondent – interviewer interactions in the model. That would be feasible when analyzing each wave of the panel survey separately. But those analyses could not model a possible evolution in the dependent variable, which is another important advantage of the joint analysis of all waves of the panel.

ACKNOWLEDGEMENTS

We thank the ISPO-PIOP Centre for Electoral Research for providing us with the data. Jacques Billiet, Marc Swyngedouw, Ann Carton and Roeland Beerten originally collected the Flemish data. The ISPO-PIOP is supported by the Federal Services for Technical, Cultural and Scientific Affairs. Neither the original collectors nor the Centre bear any responsibility for the analyses or interpretations presented here. We would also like to thank Jon Rasbash (Institute of Education, University of London) for some very useful comments about the operation of MLn. Finally we thank the referees for their constructive comments and suggestions on an earlier version of the paper.

APPENDIX 1

The rating question was: "Political parties are said to be "Catholic" or "non-Catholic". Please place the cards of the various parties on card No. 20 at the place that corresponds best to the degree in which the party is "Catholic" or "non-Catholic". If two or more parties are just as "Catholic" or just as "non-Catholic" in your opinion, place the cards on the same square. If you do not know how "Catholic" or "non-Catholic" a party might be, then simply put its card aside."

With the card:

Catholic 0 1 2 3 4 5 6 7 8 9 10 Non-Catholic

The press question was not identical for both surveys. For the first survey the press question was: "How often do you read the political news in the newspaper?"

With the response categories:

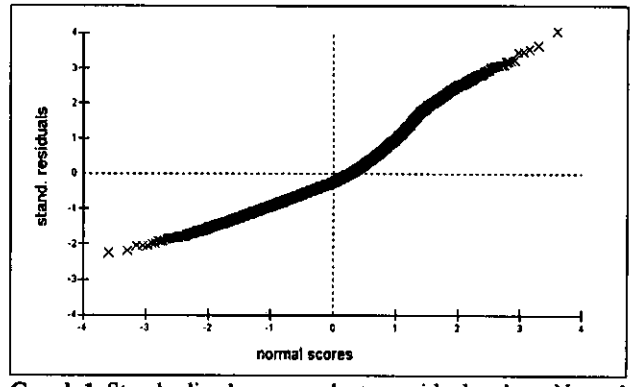
1=(almost) always, 2=often, 3=now and then, 4=seldom, 5=never

In the second survey it became: "How often do you follow the political news on the radio, on television or in the paper?"

The response categories remained the same.

APPENDIX 2

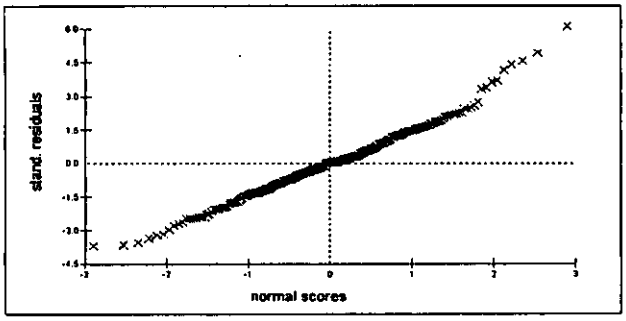
Section 4 set the assumptions of the different models that were used. For the last model the most important assumptions concern the random effects associated with the respondent and the interviewer. The assumption that the $\sigma^2_{constant}$ values for the respondent and for the interviewer are normally distributed can be assessed by looking at Normal probability plots for the residuals. Graph 1 presents the plot for the standardized respondent residuals and graph 2 the plot for the standardized interviewer residuals.



Graph 1. Standardized respondent residuals by Normal equivalent scores

In this graph the departures from the diagonal are rather limited and no apparent violation of Normality can be inferred. On the other hand it is worth noting that this graph shows more observations at the upper right hand than at the lower left end.

Graph 2 does not show any clear departures from the diagonal either. But in this graph some outliers draw the attention. Especially the outlier at the upper right hand side of the graph seems to be outside the range of the other interviewer residuals. Moreover in this graph also there are more observations at the upper right hand side than at the lower left end.



Graph 2. Standardized interviewer residuals by Normal equivalent scores.

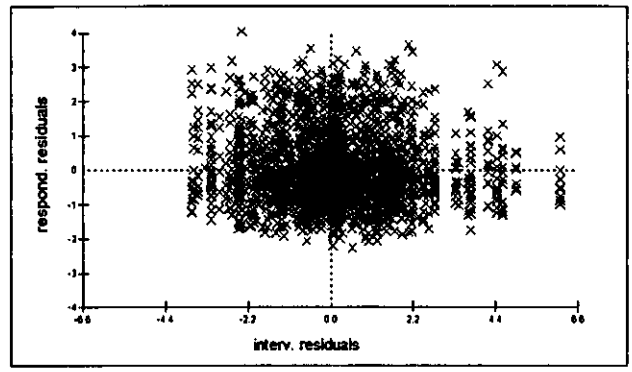
The conclusions from these graphs are as follows: there is nothing clearly wrong with the residuals but the more numerous deviations upwards and the outliers of the interviewer residuals could possibly be further investigated. Efficient techniques for these checks are not yet available for multilevel models (Goldstein 1995 29). But it is of course possible to analyze a dataset without the outliers. That is done in Table 4.

Table 4
Analysis of the Dataset Without the Interviewer Outliers (s.e. in brackets)

Fixed		
Measurement level		
constant	3.853	(0.162)
Respondent level		
sex	1.820	(0.153)
education	-1.929	(0.149)
press1	1.160	(0.090)
press2	1.217	(0.102)
Random		
Level 2		
Interviewer		
$\sigma^2_{constant}$	2.495	(0.333)
Respondent		
$\sigma^2_{constant}$	7.109	(0.530)
Measurement level		
σ^2_e	13.420	(0.481)
-2 LL		26850.2

For the analysis in Table 4 we excluded two interviewers, the one with the lowest and the one with the highest residual. The coefficients in this table are very similar to those of model c in Table 3. The interviewer variance has decreased a bit, as a result of the exclusion of extremes, but there is no evidence of a considerable impact of the outliers on the results.

The other assumption about the interviewer and respondent random effects is their mutual independence. The interviewer and respondent residuals should not correlate. That is of course more difficult to evaluate since both residuals are connected to their respective units, which do not correspond. You will get 3,028 respondent residuals and 275 interviewer residuals. An indirect check of this assumption is possible by attributing the interviewer residuals to the respondents. This is done in graph 3.



Graph 3. Standardized respondent residuals by standardized interviewer residuals

In this graph, again the more numerous deviations upwards and the interviewer outliers draw the attention. Apart from that, no pattern can be discerned. Because of the interviewer outliers, there are fewer observations at the right hand side of the graph. But the respondent residuals do not really tend to be smaller if the interviewer residuals are higher. Neither is there any evidence of the opposite.

The check in graph 3 is imperfect as it attributes the interviewer residuals to the respondents. A better alternative might be to fit a more complex model with an interaction term between the two random effects. Goldstein (1995, 119) proposes this model. A test for the model improvement due to the interaction term can give an indication for the presence of a correlation between the residuals. Another alternative is the insertion of an additional level (the region) above interviewers and respondents. That model would include a term for the regional variation, which could cause a correlation between the interviewer and respondent residuals. Snijders and Bosker (1999, 159-160) describe this model. But both models require a different parameterization with various sets of dummies. Their clarification calls for a paper in itself and is consequently outside the scope of this paper.

REFERENCES

BAILAR, B., BAILEY, L., and STEVENS, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.

BEERTEN, R., BILLIET, J., CARTON, A., and SWYNGEDOUW, M. (1997). *1995 General Election Study Flanders-Belgium. Codebook and Questionnaire*. Leuven: ISPO/Departement Sociologie, K.U. Leuven.

BRYK, A.S., and RAUDENBUSH, S. (1992). *Hierarchical Linear Models Applications and Data Analysis Methods*. Newbury Park - London: Sage.

CARTON, A., SWYNGEDOUW, M., BILLIET, J., and BEERTEN, R. (1993). *Source Book of the Voters' Study in Connection with the 1991 General Election.* Leuven: Sociologisch Onderzoeksinstituut/ ISPO.

DIPRETE, T.A., and FORRISTAL, J.D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331-357.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models.* London: Edward Arnold.

GROVES, R. M. (1989). *Survey Error and Survey Costs.* New York: Wiley.

HANSON, R.H., and MARKS, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

HOX, J.J., DE LEEUW, E.D., and KREFT, I.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys.* (Ed. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman). New York: Wiley, 439-461.

ISPO/PIOP (1995). *1991 General Election Study Belgium. Codebook and Questionnaire.* Leuven: ISPO.

KREFT, I.G., and DE LEEUW, J. (1998). *Introducing Multilevel Modeling.* London: Sage Publications.

KROSNICK, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

LOOSVELDT, G., and CARTON, A. (1997). Evaluation of nonresponse in the Belgian Election Panel Study '91 - '95. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 1017-1022.

RASBASH, J., and GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, 19, 337-350.

RASBASH, J., and WOODHOUSE, G. (1996). *MLn Command Reference.* Version 1.0a. London: Multilevel Models Project. Institute of Education, University of London.

RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

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

SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* New York: Academic Press.

SCHWARZ, N., and SUDMAN, S. (1995). *Answering Questions.* San Francisco: Jossey-Bass.

SNIJDERS, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30, 405-426.

SNIJDERS, T., and BOSKER, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

SUDMAN, S., and BRADBURN, N. (1974). *Response Effects in Surveys.* Chicago: Aldine.

YANG, M., and GOLDSTEIN, H. (1996). Multilevel models for longitudinal data. In *Analysis of Change. Advanced Techniques in Panel Data Analysis.* (Ed. U. Engel, and J. Reinecke). Berlijn - New York: Walter de Gruiter, 191-220.

Survey Methodology, December 2000
Vol. 26, No. 2, pp. 199-207
Statistics Canada

Screen Design and Question Order in a CAI Instrument Results From a Usability Field Experiment

MAREK FUCHS[1]

ABSTRACT

Screen design and questionnaire design affect the interviewer behavior in a CAI environment. Previous research has shown that interviewers can work more properly and efficiently if suitable functions and features are incorporated in the CAI instrument. Usability experiments with the household roster of two large government surveys have shown that using grids and tables is an important feature to facilitate the interviewer's performance. While these experiments were conducted under laboratory conditions, we have results from a first field experiment. In March of 1998 a CATI survey on immigrants was fielded in Germany (response rate 84%, $n = 501$). Four different versions of a household roster were compared in this production study, testing two different screen designs together with two different question orders in a 2×2 factor design. The four versions were randomly assigned to interviewers and respondents. Time measures were built into the CATI program, and 234 randomly selected interviews were video taped and analyzed according to a coding scheme. Based on the data we assessed the usability of different CAI design features. The results show that the screen design as well as the question order have a significant influence on interview duration and interviewer behaviors. Especially the grid based and topic based version allows the fastest performance in terms of time used to complete the instrument. Results from the coding data suggest that the differences between versions are due to specific interviewer and respondent behaviors. The data indicates that the grid based topic version enables a respondent oriented interviewer behavior, and thus allows the best interviewer performance in terms of duration.

KEY WORDS: Computer assisted interviewing; Usability Testing; Field experiment; Screen design; Question order.

1. INTRODUCTION

Computer assisted interviewing is on its way to becoming a standard survey technique (Couper, Baker, Bethlehem, Clark, Martin, Nicholls and O'Reilly 1998). In telephone surveys as well as with personal interviews, more and more studies are conducted using computer assisted interviewing techniques (CAI). Many of the large government surveys in the US are in the transition to CAI or have completed it already. Even in Europe, we observe a shift towards computer assisted interviewing (Schneid 1991; Fuchs 1994, 1995; Laurie and Moon 1997; Projektgruppe SOEP 1998) – even though, the methodological aspects of this development do not constitute the main focus of European research, so far.

Researchers and people responsible for fielding surveys rely on computer assisted interviewing for several reasons: (Sometimes it seems, however, that substantial arguments are less important than just a specific market rush towards CAI.)

- They hope to collect data of higher quality due to built-in consistency checks and range checks during the course of the interview.
- CAI provides the possibility to use automated skip patterns and allows to design more complex instruments without putting too much burden onto the interviewers.
- They hope to spend less time and money for interviewing and post-processing and decrease survey budgets once the up-front investment for hardware and software is payed off.
- They hope to benefit from CAI's ability to read external data into the interview which is especially interesting with panel studies.

The general movement towards CAI is evaluated positively. Researchers and field directors benefit from it (Nicholls and deLeeuw 1996) and interviewers (Couper and Burt 1994) as well as respondents (Baker 1992), reveal a great deal of sympathy or at least acceptance. On the other hand, computer assisted interviewing has introduced some additional problems into the interview situation, too: in the early years methodological research was mainly concerned with hardware and software problems (see Couper, Groves and Kosary 1989; Weeks, 1992 for overviews). Instead, recent studies dealt with interview and respondent acceptance, interview duration, and usability issues (Couper *et al.* 1998 for an overview). The present paper contributes to this later discussion of "technology effects" (Fuchs, Couper and Hansen 2000).

2. THEORETICAL BACKGROUND

For the purpose of the following analysis the theoretical focus is mainly on two usability issues: (1) segmentation of the interview flow and (2) lack of interviewer flexibility.

[1] Dr. Marek Fuchs, Catholic University of Eichstätt, Department of Sociology, Ostenstrasse 26, 85071 Eichstätt, Germany. E-mail: marek.fuchs@ku-eichstaett.de.

1. Segmentation: in a CAPI environment the interviewer has an additional burden: the process of keying takes place in the interview situation. Usually, an interviewer reads a question, receives an answer, enters the data, presses [enter] and then the next screen with the following question appears. Compared to PAPI interviewers cannot look ahead and anticipate the next upcoming question while recording the answers to the previous one and they cannot start reading the next question before pressing [enter] – they cannot work simultaneously on both tasks. As a result of this procedure the interviewer respondent interaction is segmented by [enter] keys. So far we do not have quantitative evidence that this kind of segmentation harms the data or the interview situation. But it is argued that the interviewer loses the "big picture", and the relevance of questions and their relationship to each other may be unclear (House 1985; Groves and Mathiowetz 1984).

Our findings from several series of usability tests in the lab concerning the screen layout of a household roster (Couper *et al.* 1997; Hansen, Couper and Fuchs 1998) led to the suggestion of a specific screen layout that allows the interviewer to develop a more complex understanding of the instrument, maintain the interaction with the respondent, and enter data at the same time: Two different versions of a series of questions were tested under laboratory conditions in terms of the time necessary in order to complete the questions and ease of use. We compared a so-called item based design with a grid based design. House and Nicholls (1988) distinguished between three approaches in screen design for computer assisted instruments: item based, screen based and form based design. In the item based approach one question and one input field are displayed at a time, and logic operations are performed in the transition from one item based screen to the next. This design is easy to program and focuses the interviewer's attention on the actual question. The screen based approach combines several items that need to be answered in sequence. All logic operations are executed after each item. On a form based screen, many items are presented at the same time in a table or grid and the interviewer may use the cursor keys to move from field to field and to complete them in any order.

The item version tested in our experiment matches the characteristics specified by House and Nicholls (1988) for a screen based approach. In contrast, the grid based design is best described as a form based instrument. It allows interviewers to record the information in the order chosen by the respondent, it provides the interviewer with a better overview of the instrument and it more easily allows updates and backups (for details see Couper *et al.* 1997). Also, the design matches the interviewers' demand for more questions on one screen – both for speed of administration and for context knowledge. The following graph gives an impression of an item based and a grid based CAI screen design.

We found evidence that the grid based design reduces the segmentation: interviewers could start reading the next upcoming question while still entering the data to the previous question. Even backing up seems to be easier within a grid design. On the other hand, we found only modest support for a grid based design in terms of time used to complete the task (for details see Couper *et al.* 1997). This leads to the question: what can we do to decrease segmentation and to further improve the efficiency of a household roster in terms of duration?

2. Lack of flexibility: The second feature that might cause problems in a computer assisted interview is the lack of flexibility. One of the advantages of a CAI instrument is the fact that an interviewer can hardly skip any questions. Although CAI instruments can make extensive use of skip patterns and filters, they apply a predefined question order. Usually, each question needs an [enter] key before the system goes on to the next screen. It is seen as an advantage that this rigid question order avoids any trouble the interviewer might have with the routing through the instrument, questions for specific respondents, filters and skip patterns and so on. He or she can abandon this task and focus on the administration of the actual items. On the other hand, this causes a very strict question order and provides the interviewer with little flexibility in terms of question order. A small example demonstrates this effect: most CAI instruments apply a question order to their household roster, where all items for one person are asked before the interviewer works through the same items for the next person ("person based design" see Couper *et al.* 1997; Fuchs 2001 or "grouped questions" see Moyer 1996). The CAI instrument, for example, might request the respondent's age, educational level, and other questions first before asking for the age of the respondent's wife. (This can be explained in part by the way computer programs and data bases work: households represent the main records and persons or other entities are treated as subrecords.) When completing the questions of a household roster it might happen (and in fact it happens quite often, see below) that the respondent provides not only the answer to the current question (*e.g.* "I'm 34 years old") but also to a related question: "I'm 34 years old and my wife is 32 years old" or the respondent might answer "We are all Black" when asked about his or her own race (Oksenberg, Beebe, Blixt and Cannell 1992).

While working with a paper instrument it is an easy task for an interviewer to make immediate use of the additional information provided by the respondent. In case he or she answers, for instance, "We are all black" the interviewer can easily mark the appropriate check boxes for all household members at once. For someone interested in questionnaire design this leads to the following question: given the lack of flexibility in a computer assisted environment, what is the best question order for collecting information about all household members?

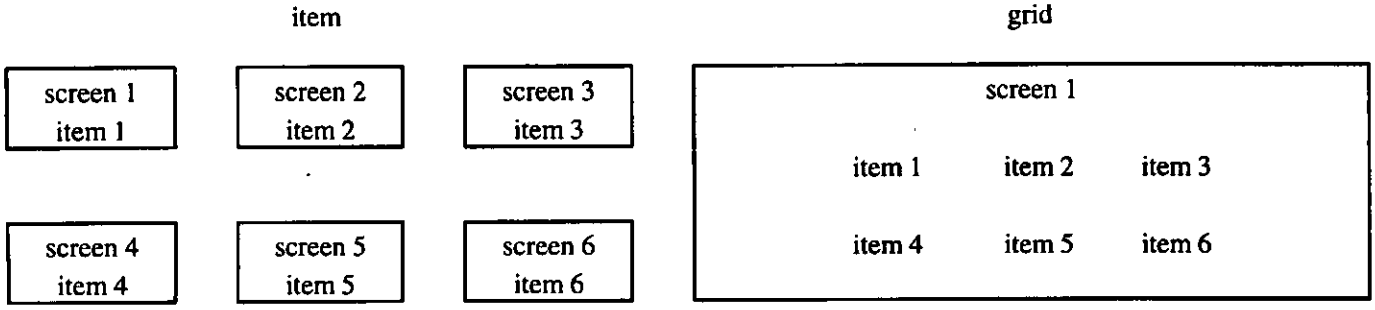


Figure 1. Item Based Design *vs.* Grid Based Design

Moore and Moyer report results from an experiment on two different question orders designed for collecting information about all eligible persons in a household (Moore and Moyer 1998a, 1998b). The first question order asks all questions for the first eligible person in the household and moves on to the next person, when all questions are completed. This question order is called a person based approach. In the second version, the topic based approach, the first question is asked for all eligible persons, then the second question for all persons and so on. Moore's and Moyer's results show strong support for a topic based design: the topic version leads to less item non-response, less break offs and refusals and is substantially shorter. Besides interviewers show significant preference for this version.

In the experiment presented in this paper we tried to make use of the advantages of a topic based approach and of a grid based screen design: we combined the two screen designs (item based design *vs.* grid based design) with the two question orders (person based order *vs.* topic based order) and tested all four resulting versions in a field experiment. In doing this, we had the following assumption in mind: the usability of a CAI instrument is not only a programming issue, but it is also connected to the questionnaire design and to the interview as a social situation. Both aspects of a computer assisted instrument, its screen design and its question order, support or hinder a smoothness of the interview flow. Based on the results of the previous research we had the following hypothesis: The combination of a grid based screen design and a topic based question order allows the most efficient interviewer respondent interaction.

3. METHODS

The experiment took place in Germany in March 1998. Immigrants of German origin from Poland, Rumania and the former Soviet Union were surveyed. Starting February 28, 1998 and ending March 20, 1998 15 interviewers completed $n = 501$ interviews. All respondents received an advanced letter and were called by phone up to 15 times. The response rate reached 84% and item non-response was considerably low. The interviews were conducted using the CATI program CI3. About 95 questions on various topics were asked. The average interview lasted 23 minutes.

Four versions of a small household roster with three items per person were included in the instrument: an item/person version, a grid/person version, an item/topic version and a grid/topic version. All versions applied the same question wording and interviewer instructions, however, we modified the screen design and the question order according to the theoretical approach mentioned before (Figure 2). The item based person version is considered to be the standard version – it represents the questionnaire design usually applied to socio-demographic portions in CAI surveys. One of the four versions was randomly assigned to each interview – and thus to interviewers and respondents. We measured the total time needed for the household roster and in addition the time spent on each single item in that section of all 501 interviews. In addition, 234 interviews were selected at random and the interviewer working through the household roster section was videotaped. The video segments were coded in terms of interviewer behavior and respondent behavior and the resulting data was combined with the time measurements.

4. RESULTS

The durations of the four versions differ significantly from each other: interviewers needed 6.6 seconds per item in the item based person version (which is considered to be the standard one). In contrast each item took 5.5 seconds in the grid based topic version. This is a reduction of about 17% for the grid based topic version. The two other versions are in between.

It is important to mention that both factors seem to contribute to the decrease in time used to complete the task. If we distinguish between the two factors, we end up with the following results: the two topic based versions are significantly shorter than the two person based versions and the two grid based versions take significantly less time than the two person based versions. The combined effect applies to

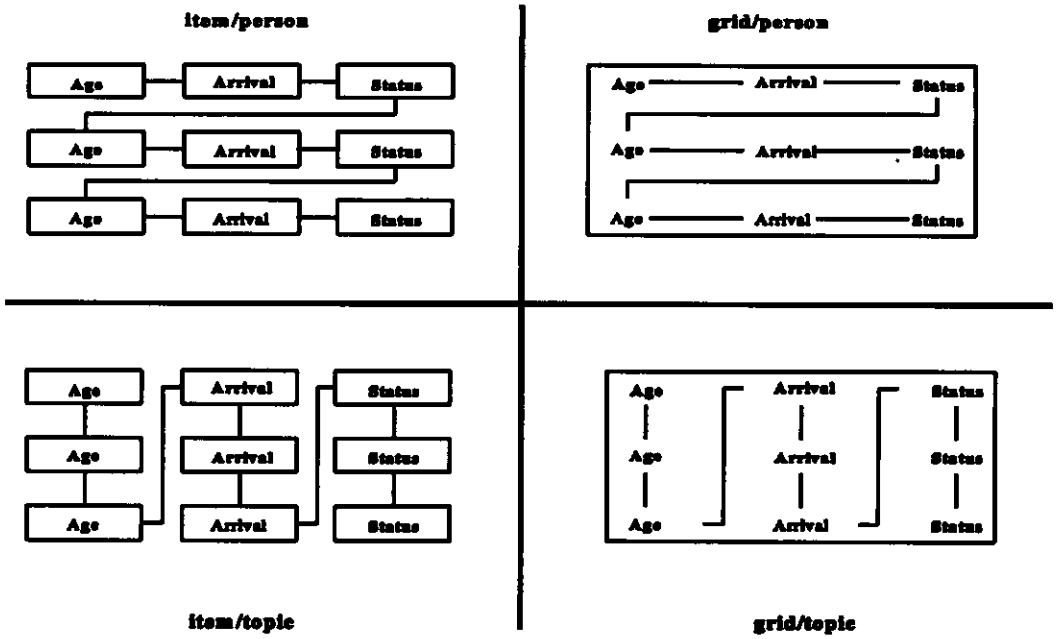


Figure 2. Four Versions Tested in the Experiment (Each Box Represents One Screen)

the grid based topic version and leads to the value of 5.5 seconds per item. (An analysis of variance reveals that both factors – the screen design as well as the question order – contribute independently to the decrease in time (screen design: $p<0.01$, one third of total effect; question order: $p<0.001$, two thirds of total effect, no significant interaction).)

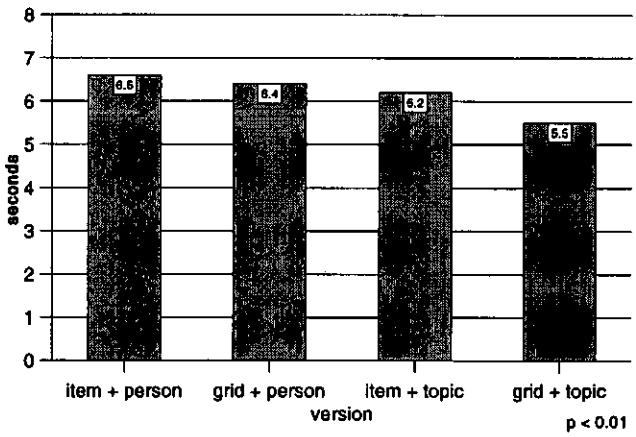


Figure 3. Duration per Item by Version

But why is the grid based topic version faster? A detailed analysis shows that this version is especially faster when collecting the information for the second and all following persons in the household – a significant impact, that is called loop effect (Fuchs 2001). This term describes the following phenomenon: the interviewer takes much longer to collect the information for the first person in a household compared to all subsequent persons. The average loop effect sums up to 3.4 seconds per item which is a reduction of about 38% compared to the first person (Table 1).

The loop effect is not specifically characteristic for this experiment. We recognized loop effects in our previous experiments with the NHIS household roster, too (Couper *et al.* 1997). It is, however, interesting to observe that the loop effect is significantly larger for the topic based versions than it is for the versions that follow a person based question order (Table 1). Thus the topic based versions do increase the acceleration for the second and all subsequent persons in a household and consequently show a larger loop effect. (One implication of our experimental design might be that interviewers did not know what version they were approaching. This may have decreased their performance on the very first item. But this effect should be the same across all versions, so the results should not be affected.)

Table 1
Duration and Loop Effect (Seconds)

Item	Age	Arrival	Status	All items
	Duration per item			
First person in household	9.4	9.9	7.7	9.0
All other persons in household	6.6	5.7	4.5	5.6
All persons	8.0***	7.8***	6.1***	7.3***
	Loop effect			
Differenz between first and all other persons in the household	-2.8	-4.2	-3.2	-3.4
	Loop effect by version			
Grid + topic	-6.8	-6.4	-4.6	-5.9
Item + topic	-5.2	-8.3	-4.3	-6.0
Grid + person	-0.5	-2.7	-3.0	-2.1
Item + person	0.3	-0.3	-1.3	-0.4
Average loop effect	-2.8***	-4.2**	-3.2**	-3.4***

**p < 0.01; ***p< 0.001

Analyzing the video tapes we can provide reasons for these differences at least in part: given the topic based conditions, both interviewers and respondents adapt differently to the interview situation compared to the person based versions. When asking the questions for all persons in the household, the respondent recognizes the logic of the procedure very quickly. In quite a high proportion of all cases (about 30%) their reaction to this is "We all arrived in the same year" (meaning: "Don't ask me this question again and again").

If the instrument follows a person based design, the interviewer has to memorize this piece of information, and if it comes to the next person, he or she needs to remember: "Do not ask this question again, the respondent gave you the appropriate answer already!" Only in a few cases they really do, most of the time they just ask the question again. This is especially true when using an item based screen layout that gives no clues in terms of the answers to the same question for the other household members. In a topic based design instead, the interviewer can easily adapt to that situation. Thus he or she just enters the same code for all persons in the household without asking the question repeatedly. Both the interviewer and the respondent get used to the questions, and so the question answer process runs with less verbal contributions from the interviewer's side as well as from the respondent's. Both interviewer and respondent can anticipate the next question and the interview runs more smoothly. This is especially true when the CAI instrument makes use of a grid and provides further context information, *e.g.*, the responses for other household members to the same question. (Looking at the results reported in the lower part of Table 1 we conclude that the grid based person version does not benefit to the same extent from the advantages of the topic based approach. However, due to the grid design the loop-effect is considerably larger than in the item based person version.) As a result the time used per item is substantially shorter and the interviewer can provide respondent oriented interviewer behavior similar to Schober and Conrad's (1997) findings.

Providing feedback by the interviewer sometimes works as a signal that he or she has recorded the answer to the previous question in order to stimulate the respondent, so that the latter guesses about the next question and reveals the appropriate answer even without an additional stimulus. In extreme this might lead to a respondent behavior where he or she provides the information about all persons in the household at once: "We all came in the same year". The different versions tested in this experiment impel and support such behaviors to different degrees. From our results we can conclude that the grid based topic version stimulates interviewers and respondents to deviate from the scripted interview to a higher degree than the other versions. As far as duration is concerned this version allows the interviewer to make efficient use of information provided for all household members at once. Evidence from the video coding support our interpretation of version-specific occurrences of time saving interviewer behaviors (1) and respondent behaviors (2):

1. By means of analyzing the video tapes we observe quite a lot of interviewer behaviors that do not follow standard interviewer procedures: besides the fact that about 78% of all items are read as worded, interviewers do not administer 9.3% of all items to the respondent. In another 5% of instances, the interviewer does not read the question but instead provides a different stimulus containing the relationship of the next person to the respondent (*e.g.*, "... and your wife?"). (It is interesting to recognize that interviewers chose the same verbal expressions on their own that Moore and Moyer 1998a, 1998b scripted in their experiments on question order.) In 5.5% of all cases the interviewer does not read the question but rather verifies the answer ("... and your wife is 32 years old?"). Some incomplete questions and wrong fills are observed, too. In total we have about 22% of all items affected by at least one interviewer behavior that does not follow a standardized interview script – which is a surprisingly high value considering that all interviewers were aware of the fact the interviews were video taped! Compared to other studies on interviewer behavior, however, the values are considerable lower. For example Oksenberg, Cannell and Blixt (1996) applied behavior coding to the National Medical Expenditure Survey and reported 37% to 41% of such interviewer behaviors. We will come back to the question of whether or not these behaviors help obtain valid measurements.

We draw the following conclusion from these particular findings: most of these behaviors indicate kind of a short-cut, *e.g.*, the interviewer does not read the question text as worded, he or she tries to make the conversation smoother and more suitable in terms of conversational rules. From our point of view this indicates that interviewers do not want to ask for information the respondent provided already. They do not want to behave unresponsively toward the verbal contributions of the respondent, instead, they wish to follow conversational rules. As a side effect these behaviors are less time consuming than standard interviewer behaviors. In our perspective, the priority therefore lies not with saving time, but with customizing the question answer process to respondent behaviors not anticipated and not absorbable by the computer assisted instrument.

In order to compare the four screen design versions in terms of the degree of interviewer deviations from the standard interview script we have computed the proportion of items per case affected by this kind of behavior. Large differences in interviewers not following the scripted interview between the four versions are to be noticed: Applying the grid based topic version to an interview results in more than twice as many such behaviors (the average proportion

of items affected is 0.48) than the item based person version (0.21) which is the standard for most studies so far. (An analysis of variance indicates that both factors contribute independently to the overall effect (screen design: $p < 0.001$; question order: $p < 0.001$; no significant interaction effect). About 25% of the overall effect can be attributed to the screen design, about three quarter to question order.) And this contributes to the time used for interviewing: items affected by a interviewer behavior not scripted in the interview take substantially less time (4.0 seconds) than the regularly administered items (6.8 seconds; $p < 0.001$).

Table 2
Interviewer Behavior and Respondent Behavior by Version

	Grid + topic	Item + topic	Grid + person	Item + person	Total
Average proportion of items affected by interviewer behavior not following the scripted interview per case	0.48	0.43	0.34	0.21	0.36***
Respondent provides information for all persons in the household at once	38.2%	44.4%	29.0%	10.8%	29.7%***

***$p < 0.001$
In order to differentiate between the proportion of cases affected by a certain respondent behavior and the average proportion of items per case (!) affected by a certain interviewer behavior we used percent notation for the first and decimal notation for the later.

2. Additionally an analysis of the respondents' behavior shows that the topic design leads to a higher proportion of cases (42,3% compared to 19.7% for the person approach; $p < 0{,}001$) where the respondent provides at least once in the household roster section the information for all persons or a group of persons at once (*e.g.*, "We all came in the same year"; "We all have the same legal status"). By contrast, the difference of the grid based design from the item based design is considerably smaller (33,6% *vs.* 26.1%) but does not reach the level of significance. However, an analysis of the interaction reveals a significant interaction effect ($p < 0.05$): Using a topic oriented question order the grid design does not make a significant difference. However, on top of an topic oriented question order the grid design increases the number of instances where the respondent provides the information for all household members at once.

It is surprising that results differ even for the two screen designs when using a person oriented question order. The study was administered by telephone, the respondents not being aware of the screen design at all. The only possible explanation is based on the fact that the interviewers modify their behavior in concordance with the screen design, stimulating the respondent differently. Accordingly, respondents, as well as interviewers, react to the screen design and the question order under the grid based person design in a way that facilitates the interviewer respondent interaction and thus helps smoothen the interview flow. (As seen before, the interviewers change their behavior even under the grid based person condition (Table 2), however, the question order does not stimulate respondents to behave accordingly.)

One possible drawback of these interviewer and respondent behaviors might be a lack of data quality due to changes occurring in the predefined question answer process; instead, the respondent considers the answer less intensively and thoroughly. We observe only very few item missing values so an analysis of this standard indicator for data quality is not efficient. In fact, we do not expect a higher proportion of item missing values in either version. One might, however, be concerned about the homogeneity of the answers provided by the respondent. In a high proportion of cases he or she listens to the full question text only once and that could contribute to a less thorough consideration when answering the same question for subsequent household members. Additionally, answering for all household members at once ("We all arrived in the same year") might increase the homogeneity of the response and thus decrease data quality.

Table 3
Average Number of Different Categories (Homogeneity) per Household by Version

Variable	Grid + topic	Item + topic	Grid + person	Item + person	Total
Year of arrival (19 categories)	1.2	1.2	1.2	1.3	1.2
Status (4 categories)	1.3	1.3	1.3	1.3	1.3

No significant differences

In order to assess this possible drawback we computed the number of different response categories chosen by the respondent on a particular item for all household members (*e.g.*, for year of arrival: respondent 1985, partner 1987, daughter 1987, son 1988 = 3 different response categories). This should give us an idea of whether or not only those respondent make use of the short-cut ("We came all in the same year") for whom this is actually valid, or whether even other respondents provided one answer for all household members even though they should have chosen two or more different response categories because of the situation in their particular household (unfortunately we have no external validation for the responses provided). In looking at the average number of different response categories (Table 3) we do not notice any differences in terms of homogeneity of data. For the year of arrival as well as for the legal status (as a German or a foreigner) there is no visible difference between the versions. For all versions the average number of different response categories chosen (one for each person in a household) shows no significant difference.

These finding provide only weak evidence that a grid design does not harm data quality. Other standard data quality indicators need to be assessed with larger data sets

in order to decide whether or not data quality is affected. However, based on the data available, we are unable to proof a effect on the validity of the responses.

5. DISCUSSION AND CONCLUSION

Our results from a comparison of four versions for a household roster (using the same question wording across versions) indicate that interviewers as well as respondents perform more efficiently under the grid based topic condition than with the other three versions. Combining a grid based screen design and a topic based question order reduces the average duration by about 17%. Two thirds of this reduction can be attributed to the question order, approximately one third to the screen layout. It is important to mention that the effect of the screen design is less pronounced than the one of question order and – compared to the effect on duration – even smaller on interviewer behavior and respondent behavior.

Even though the effects of the grid design on interviewer behavior and respondent behavior are far from large, they help to elicited two reasons for the better performance of the grid based topic version in terms of interview duration: (1) in the grid based topic version, the interviewer as well as the respondent adapt better to the logic of the question answer process, both anticipate the next question more easily and the question answer process runs more smoothly. (2) This version leads to more occurrences in which the respondent provides the information for the persons in the household faster and more often the respondents reveal the information for all household members or at least for one group at once. Even though the results are not fully consistent, this particular version makes it easier for the interviewer to adapt to this situation, record the information and stimulate the respondent to give the next appropriate answer without repeating the full question text.

Our findings contribute to the discussion of how to design survey instruments for interviewer administered computer assisted data collection. Based on the results reported in this paper we can draw the conclusion that making use of grids facilitates the interviewer respondent interaction and helps speed up data collection. Our experiments on item design *vs.* grid design conducted in the University of Michigan Survey Research Center's usability laboratory have shown that we can improve interviewer performance by providing grids (Couper *et al.* 1997). Moore and Moyer (1998a; 1998b) have demonstrated that one can improve interview efficiency by switching to a topic based question order, too. The present paper indicates that the interview situation benefit even more when combining both features.

Using grids and a topic based question order causes a greater amount of instances where the interviewer deviates from the scripted interview. From a rigid methodological point of view this might be seen as an important drawback, especially, if the interviewer deviates from the standard interview script using global questions for all household members. For example, Martin (1999) showed a significant increase in the number of people enumerated in a household if extra questions were asked. In addition, Kindermann and colleagues (1997) demonstrated for non-household roster type questions, that additional cues on victimization significantly increase enumerations of crimes. Generalizing these results to global questions across persons, one would expect a decrease of data quality as interviewers use non-scripted behaviors that apply global questioning methods. However, the results reported in this article do not indicate that interviewers are using global questions and the author does not recommend to make extensive use of global questions when designing a survey instrument. Instead, the findings lead to a screen design that allows interviewers to make use of information reported when the respondent switches to a global mode and provides the information for several household members at a time. So, we do not want to encourage researcher to make extensive use of global questions and we do not want to see interviewer modify the scripted questions in order to ask global ones. However, when confronted with a respondent providing more information than actually asked for, the screen design of the CAI instrument should not prevent interviewers from making use of it.

A grid based design has been proven to facilitate the interviewers job with respect to this task, because it allows interviewers to adjust their behaviors in concordance with general conversational rules. Basic findings of behavior coding suggest that interviewers frequently deviate from specific interviewing procedures. "These changes often reflect adjustments made by the interviewers to meet the exigencies of the situation: to melt it more congenially with communications immediately preceding it, or to adjust to the respondent's particular situation" (Oksenberg *et al.* 1992: 3). This is especially necessary when respondents do not limit their answers to the information requested by the question, but elaborate it or provide additional information. "Avoiding the appearance of not paying attention to the respondent, interviewers in this situation frequently filled in the answer themselves without asking the question, or asking it only in part" (Oksenberg *et al.* 1992: 5). They thus try to switch to more respondent oriented procedures to avoid looking unresponsive. A grid based screen design and a topic oriented question order supports interviewers to interact according to these conversational rules and with respect to the interview situation's needs. This might be acceptable or even preferable as long as we are talking about factoid questions and as long as these interviewer behaviors do not harm data quality (*e.g.*, leading question or probes).

What needs to be done in order to improve the computer assisted instrument in its supporting function for the interviewer respondent interaction: Our data suggests that the grid based topic version leads to a specific interview flow,

so that interviewer and respondent can easily adapt to it. Jeff Moore (1996; Moore and Moyer 1998a, 1998b) has shown that interviewers prefer the topic based version. By contrast, we know little about the respondents' satisfaction with that question order. Assessing their opinion about the different version is consequently an important goal. Moreover, we do not know whether this version matches the way in which information is stored in the respondents' brains. It might be, that respondents can easily adapt to this version, but that in terms of cognitive and social burden or in terms of correctness of answers it is not the right method. Additionally, we need to focus on the question whether or not we can transfer our findings from a household roster to other segments of a questionnaire. Right now we are conducting a series of field tests comparing different design solutions for factoid information other than household roster information and for attitude items. The versions tested in this experiment differ in the degree of contextual information provided to the interviewer while administering a particular item (previous questions, next questions *etc.*). The general question sounds: what happens if we use grids or form based screens more extensively? Under what conditions and circumstances does it help to improve interview efficiency and what are the limitations to this approach? However, it is too early to present any results at this time.

In addition, there are more unanswered questions that need to be addressed in future research. Personally I would like to suggest a specific approach to assess these questions assuming that computer assisted instrument design is of importance to different clients: researchers, interviewers and respondents. Of course, it is important that a CAI instrument meets the researcher's needs to obtain his or her measurements and also that the question answer process be well designed for each single item. However, in my view considering the social dimension of the interviewer respondent interaction and the behaviors in between single items is also a matter of importance. If the CAI instrument disturbs the social dimension of the measurement process it might harm even data quality. So far we do not know which approach allows the best compromise between validity and reliability of the measurement process on the one hand and a smooth short and non-embarrassing interview flow on the other hand. In order to find out to what respect a specific CAI screen design might harm data quality and how it helps save time, money and interviewer effort we need to conduct more usability studies.

To assess the questions mentioned above we do need more field experiments. Due to the fact that we want to analyze the social dimension of the interview and its effects on interviewer behavior as well as on interview duration, laboratory experiments do not meet our needs completely. Of course laboratory experiments allow a more controlled setting, reveal more detailed information about both participants, and – as a result – need smaller numbers of cases. Still, without going into the field, we will never confront our prototypes and design solutions with real pressure to maintain and facilitate the interviewer respondent interaction and the question answer process at the same time. Usability testing should therefore be seen as a joint process of laboratory experiments and field tests.

ACKNOWLEDGEMENTS

Some of these results were presented at the SMP Brown Bag Seminar, Institute for Social Research, University of Michigan on May 21, 1998 and on the occasion of the 54th Annual Meetings of the American Association for Public Opinion Research, May 16, 1999. Special thanks go to the interviewers participating in this experiment. Mick Couper, Siegfried Lamnek, Jeffrey Moore and two anonymous reviewers provided helpful suggestions on earlier versions of this paper.

REFERENCES

BAKER, R.P. (1992). New technology in survey research: computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 10, 145-157.

COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z.F., MARTIN, J., NICHOLLS, W.L., and O'REILLY, J. (Eds.) (1998). *Computer Assisted Survey Information Collection.* New York: Wiley.

COUPER, M.P., and BURT, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 12, 38-54.

COUPER, M.P., GROVES, R.M., and KOSARY, C. (1989). Methodological issues in CAPI. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-354.

COUPER, M.P., FUCHS, M., HANSEN, S.E., and SPARKS, P.(1997). CAPI Instrument Design for the Consumer Expenditure (CE) Quarterly Interview Survey. Final Report. University of Michigan.

FUCHS, M. (1994) *Umfrageforschung mit Telefon und Computer.* Einführung in die computergestützte telefonische Befragung. Weinheim: Psychologie Verlags Union.

FUCHS, M. (1995). Die computergestützte telefonische Befragung. Einige Antworten auf Probleme der Umfrageforschung. *Zeitschrift für Soziologie*, 24, 284-299.

FUCHS, M. (2001). The impact of technology on interaction in computer-assisted interviews. (Ed. D. W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and H. van der Zouwen). *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview.* Wiley (forthcoming).

FUCHS, M., COUPER, M., and HANSEN, S. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics* (in press).

GROVES, R.M., and MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effect on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.

HANSEN, S.E., COUPER, M.P., and FUCHS, M. (1998). Usability Evaluation of the NHIS Instrument. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.

HOUSE, C.C. (1985). Questionnaire design with computer assisted telephone interviewing. *Journal of Official Statistics*, 1, 209-219.

HOUSE, C.C., and NICHOLLS, W.L. (1988). Questionnaire design for cati: design objectives and methods. (Ed. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg). *Telephone Survey Methodology*, New York: Wiley, 421-436.

LAURIE, H., and MOON, N. (1997). Converting to CAPI in a Longitudinal Panel Study. Working papers of the ESRC Research Centre on Micro-Social Change, 97-11, Essex.

MARTIN, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220-236.

MOORE, J.C. (1996). Person- vs. Topic-based Design for Computer-Assisted Household Survey Instruments. Paper presented at InterCASIC '96, International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX.

MOORE, J.C., and MOYER, H.L. (1998a). ACS/CATI Person-Based/Topic-Based Field Experiment – Final Report. Center for Survey Methods Research, U.S. Bureau of the Census.

MOORE, J.C., and MOYER, H.L. (1998b). Questionnaire Design Effects on Interview Outcomes. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.

MOYER, L.H. (1996). Which is better: grid listing or grouped questions design for data collection in establishment surveys? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 986-990.

NICHOLLS, W.L., and de LEEUW, E. (1996). Factors in acceptance of computer-assisted interviewing methods: a conceptual and historic review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 758-763.

OKSENBERG, L., BEEBE, T., BLIXT, S., and CANNELL, C. (1992). *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews*. Final report. Survey Research Center, Ann Arbor, USA.

OKSENBERG, L., CANNELL, C., and BLIXT, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. U.S. Department of Health and Human Services. AHCPR No. 96-N016.

PROJEKTGRUPPE SOEP (1998). Funktion und Design einer Ergänzungsstichprobe für das Sozio-oeconomische Panel. Diskussionspapiere des DIW, 163, Berlin.

SCHNEID, M. (1991). Einsatz computergestutzter Befragungssysteme in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. ZUMA-Arbeitsbericht 91/20. Mannheim: ZUMA.

SCHOBER, M.F., and CONRAD, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

SUCHMAN, L., and JORDAN, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 45-54.

WEEKS, M.F. (1992). Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. *Journal of Official Statistics*, 8, 445-465.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 2000. An asterisk indicates that the person served more than once.

J.-F. Beaumont, *Statistics Canada*
W. Bell, *U.S. Bureau of the Census*
* D.R. Bellhouse, *University of Western Ontario*
Y. Berger, *University of Southampton*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistics Canada*
D. Cantor, *Westat Inc.*
P. J. Cantwell, *U.S. Bureau of the Census*
R.G. Carter, *Statistics Canada*
J. Chen, *University of Waterloo*
C.Z.F. Clark, *U.S. Bureau of the Census*
M. Cohen, *National Center for Education Statistics*
* J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistics Canada*
* J.L. Eltinge, *U.S. Bureau of Labor Statistics*
* W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
* M.A. Hidiroglou, *Statistics Canada*
D. Holt, *Office for National Statistics, U.K.*
J.-S. Hwang, *Academia Sincia*
* D. Judkins, *Westat, Inc.*
* G. Kalton, *Westat, Inc.*
S. Kaufman, *National Center for Education Statistics*
J. Kim, *Westat Inc.*
* P.S. Kott, *National Agricultural Statistics Service*
* M. Kovačević, *Statistics Canada*
M. Kramer, *U.S. Bureau of the Census*
P. Lahiri, *University of Nebraska - Lincoln*
N. Laniel, *Statistics Canada*
* M. Latouche, *Statistics Canada*
P. Lavallée, *Statistics Canada*
S. Linacre, *Australian Bureau of Statistics*
* S. Lohr, *Arizona State University*
* H. Mantel, *Statistics Canada*
P. Merkouris, *Statistics Canada*
J. Moloney, *Statistics Canada*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Norris, *Statistics Canada*
J.-S. Pischke, *Massachusetts Instituteof Technolory*
D. Pfeffermann, *Hebrew University*
N. Plante, *L'Institut de la Statistique du Québec*
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*
K. Rust, *Westat Inc.*
I. Sande, *Telcordia Technologies*
N. Schenker, *University of California - Los Angeles*
F.J. Scheuren, *The Urban Institute*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University*
* J. Shao, *University of Wisconsin - Madison*
* M.P. Singh, *Statistics Canada*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R.T. Smith, *National Agricultural Statistics Service*
E. Stasny, *Ohio State University*
* D. Stukel, *Statistics Canada*
A. Théberge, *Statistics Canada*
* Y. Tillé, *Ecole Nationale de la Statistique et de l'Analyse de l'Information*
S. Tremblay, *Statistics Canada*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
M. Wendt, *Statistics Canada*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
* W. Yung, *Statistics Canada*
* E. Zanutto, *University of Pennsylvania*
A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 2000 issues: J. Beauseigle (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Corbeil, C. Ethier, C. Larabie, D. Lemire and G. Ray of Household Survey Methods Division, for their support with coordination, typing and copy editing.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 28, No. 2, June/juin 2000, 225-448

CONTENTS TABLE DES MATIÈRES

Volume 28, No. 3, June/juin 2000, 449-672

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 16, Number 1, 2000

Volume 16, Number 2, 2000

Volume 16, Number 3, 2000

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.