C3

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1998
•
VOLUME 24
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1998 • VOLUME 24 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

July 1998

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 $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: 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, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 24, Number 1, June 1998

CONTENTS

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Kott, Amhrein and Hicks tackle the problem of multi-purpose surveys. For such surveys, it would be desirable to be able to stratify the target population in various ways in order to improve the prècision of the estimates of interest. The authors present four sampling methods for the selection of samples through various stratifications while reducing the overall size of the sample. These strategies are then evaluated using data taken from an agriculture survey. The authors then show how a calibration estimator can improve the relative efficiency.

Singh, Horn and Yu examine the problem of estimating the variance of the general linear regression estimator. They carry out calibration at two distinct levels. The higher-level calibration thus defined uses the known total and variance of the auxiliary variables. The authors show that this method covers a broader range of estimators than the lower-level calibration method, which uses only the known total of the auxiliary variables. An empirical study is presented to assess the efficiency of the proposed strategies.

Hidiroglou and Särndal concern themselves with the use of auxiliary data in two-phase sampling. They explain how these data are converted into calibration weight, in two phases, in order to create efficient estimators of a population total. The authors show that the calibration estimator, using the generalized least squares function, can be expressed as a perfectly equivalent two-phase regression estimator, that is, an estimator that is the product of two successive regression fits. They examine forms of the two-phase calibration estimator when the auxiliary data are for population subsets known as "calibration groups." They also discuss the estimation of domains of interest and the estimation of variance.

Byczkowski, Levy and Sweeney consider survey frames having a many-to-many structure, that is, any unit in the frame may be associated with multiple target population elements and any target population element may be associated with multiple frame units. This problem is motivated by a building characteristics survey in which the target population consists of commercial buildings, but the frame consists of a list of street addresses (which in turn correspond to either single buildings, multiple buildings or parts of buildings). Under this setting, estimators of totals and means and their variances using simple and stratified random sampling without replacement are developed.

Yansaneh and Fuller present a recursive regression estimation procedure to reduce the computational complexity associated with best linear unbiased estimation in the context of a repeated survey with partial overlap. They use data from the U.S. Current Population Survey (CPS) to compare variances of their recursive regression estimator to some alternative estimators including the current CPS composite estimator. The proposed estimator seems to be very competitive for estimates of both level and change. They also estimate variances under various rotation patterns and find that the current 4-8-4 rotation pattern is superior to continuous rotation for current level and long-period averages, but inferior for short period changes.

Lehtonen and Veijanen bring together two well-known ideas, generalized regression (GREG) and pseudo maximum likelihood estimation, to develop a new methodology for estimating the population total of a categorical survey variable, given a vector of known auxiliary variables. The values of the categorical variable are modeled as realizations from a multinomial logistic and the corresponding unknown parameters are estimated through pseudo maximum likelihood. Then, the population frequencies of interest are estimated via a modified GREG estimator which uses these estimated parameters. Variance estimates of the frequencies are given through Taylor linearization, and some empirical results based on Finnish Labour Force Survey data are provided.

Casady, Dorfman and Wang consider the construction of confidence intervals for domain parameters in the case where the domain sample size is not fixed by the design. They condition on the observed domain sample size and show how, under certain assumptions about the population, conditional t-based confidence intervals can be obtained. In an empirical study using data from the U.S. Bureau of Labor Statistics Occupational Compensation survey, they demonstrate that the proposed conditional intervals have better coverage probabilities than standard marginal intervals.

Montanari compares two well-known estimators of a finite population mean: the GREG and the design-optimal regression estimator obtained from the difference estimator. While the former can be inefficient if the underlying model is misspecified, the latter, although model-free, is vulnerable to sampling fluctuations. An efficiency measure, which provides a criterion for choosing between the two estimators, is given. The results of an empirical study, which investigates the behaviour of both estimators under a variety of misspecified and correct models, are discussed.

Haines and Pollock provide a fresh examination of estimating totals with multiple frames. Estimators are developed when information is only available from list frames and, in addition, when information is also provided from an area frame. A simulation shows that the best estimator depends on the known, or assumed, dependence of the frames. They also study the situation when observations are either available for all units or only available for a sub-sample from each frame. Again, the preferred estimator changes when the dependence between frames is considered.

Bates and Gerber analyze the dynamics of a difficult problem: how temporary mobility of an individual contributes to within-household coverage error. They develop a two dimensional typology to characterize temporary mobility, then using data from the Living Situation Survey, conducted in the U.S. in 1993, they identify four temporary mobility patterns. Two of these traits are found to be useful predictors of persons missed from censuses or surveys.

The Editor

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 3-9
Statistics Canada

Sampling and Estimation From Multiple List Frames

PHILLIP S. KOTT, JOHN F. AMRHEIN and SUSAN D. HICKS[1]

ABSTRACT

Many economic and agricultural surveys are multi-purpose. It would be convenient if one could stratify the target population of such a survey in a number of different ways to satisfy a number of different purposes and then combine the samples for enumeration. We explore four different sampling methods that select similar samples across all stratifications thereby reducing the overall sample size. Data from an agriculture survey is used to evaluate the effectiveness of these alternative sampling strategies. We then show how a calibration (*i.e.*, reweighted) estimator can increase statistical efficiency by capturing what is known about the original stratum sizes in the estimation. Raking, which has been suggested in the literature for this purpose, is simply one method of calibration.

KEY WORDS: Calibration; Collocated sampling; Permanent random numbers; Poisson sampling; Systematic probability proportional to size sampling.

1. INTRODUCTION

Many of the list frame surveys conducted by the National Agricultural Statistics Service (NASS) are integrated in the sense that data on a range of heterogenous items, such as planted crop acres and grain stock inventories, are collected in a single survey rather than through a number of independent surveys. Bankier (1986), Skinner (1991), and Skinner, Holmes and Holt (1994) have shown how an old method of combining independently drawn stratified simple random samples – where each sample comes from a (list) frame with a different stratification scheme – can be made more efficient; that is, the variances resulting from such a combined estimation strategy would not be as large as those from the independent surveys summarized by themselves.

Even more appealing for many applications would be a sampling design that tends to select the same units from every frame, thereby reducing both the cost and respondent burden of an integrated survey. This paper explores several such designs. Three make use of permanent random numbers. The fourth uses a variation of systematic probability proportional to size sampling. The goal for each is to meet or exceed – at least on average – a particular set of sample size targets.

The paper shows how a calibration (*i.e.*, reweighted) estimator can provide relative efficiency by capturing what we know about the original stratum sizes in the estimation. A final section points out that the use of a calibration technique can do more than simply reflect original stratum sizes.

An alternative strategy for burden reduction is to use separate instruments for different survey targets and to select distinct samples for each instrument. This increases the number of units selected over all, but reduces the burden per selected unit. NASS is using that approach in its Agricultural Resources Management Study (see Kott and Fetter 1997), but it is *not* the approach to be discussed here.

2. INDEPENDENT SAMPLING AND UNBIASED ESTIMATION

Suppose we have F independent frames; for example, a sorghum frame, an oats frame, and a general grain stocks frame. Each frame is stratified independently, and without replacement simple random samples are drawn from each stratum of every frame. Frame f (say, the oats frame) contains H_f strata; stratum h (large oats operations) in frame f has N_{fh} population units, out of which n_{fh} units are selected. The union of the F frames must cover the entire (list) population, but no single frame need be complete. The frames may overlap.

One unbiased estimator for a population total $T = \sum_{i\in P} y_i$ is the simple multiplicity estimator suggested by Skinner (1991):

$$t_M = \sum\nolimits_{i\in P} y_i n_{(i)}/E[n_{(i)}], \tag{1}$$

here P denotes the entire population, and $n_{(i)}$ is the number of times unit i is selected for the sample from any frame. Observe that $n_{(i)} = 0$ for the population units not in the sample. In the great majority of applications, $n_{(i)}$ will be one for most sampled units, but $n_{(i)} > 1$ is a possibility with this design.

The expected number of times unit i will be selected for the sample is $E[n_{(i)}] = \sum^F p_{if}$, where p_{if} is the probability of selecting unit i in the stratified simple random sample from frame F; that is, $p_{if} = n_{fh}/N_{fh}$, where unit i is in stratum h of frame f.

There is also a Horvitz-Thompson estimator for T under the design, namely $t_{\mathrm{HT}} = \sum_{i\in S} y_i/\pi_i$, where S denotes the sample and $\pi_i = 1 - (1 - p_{i1})(1 - p_{i2})\cdots(1 - p_{iF})$. See Bankier (1986) for further discussion of this approach.

[1] Phillip S. Kott, Research Division; John F. Amrhein, Survey Sampling Branch; and Susan D. Hicks, Estimates Division, National Agricultural Statistics Service, USDA.

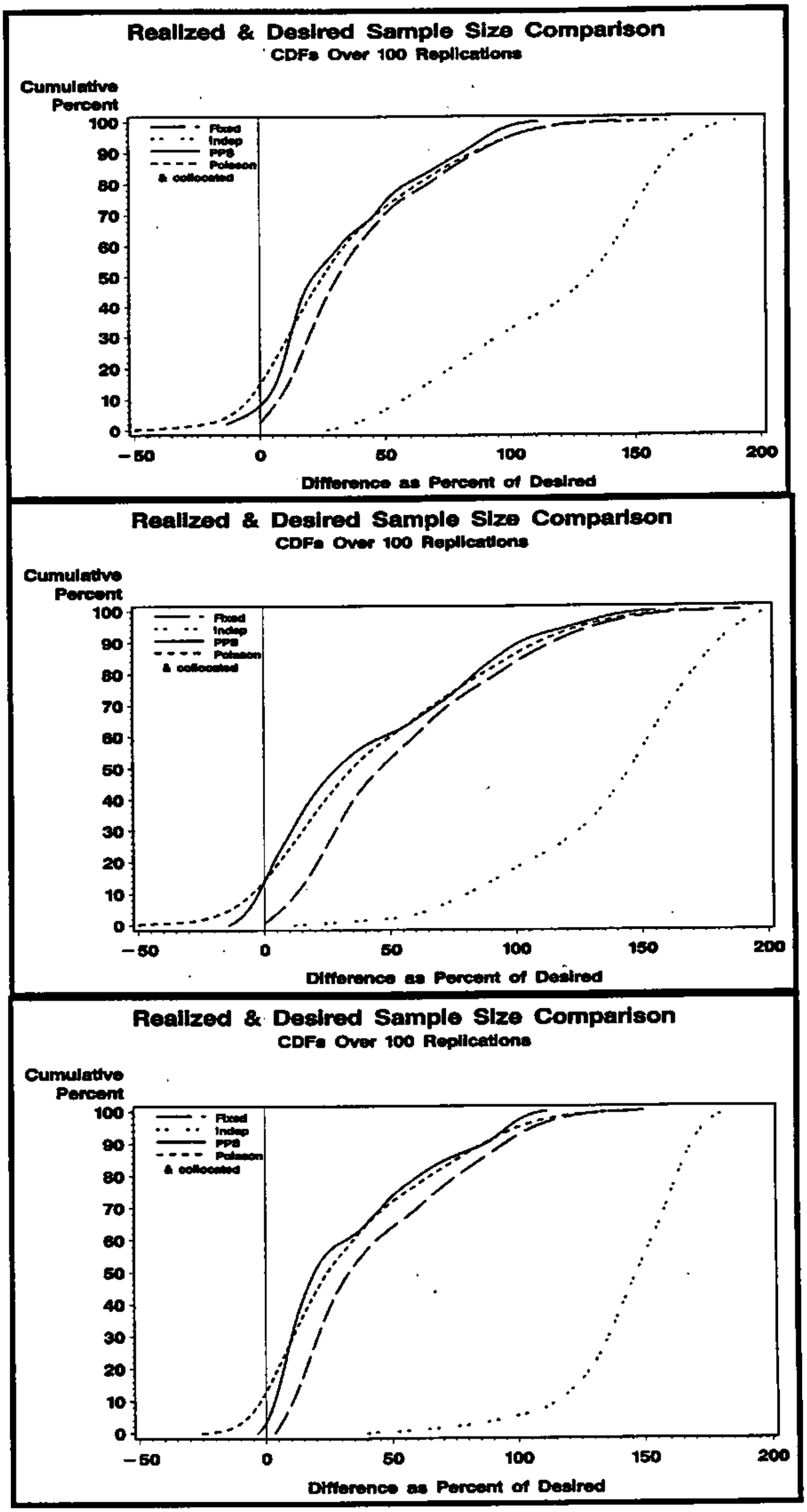


Figure 1. Comparison of realized and desried sample sizes for sampled strata. Top - MI; middle - CA; bottom - NJ.

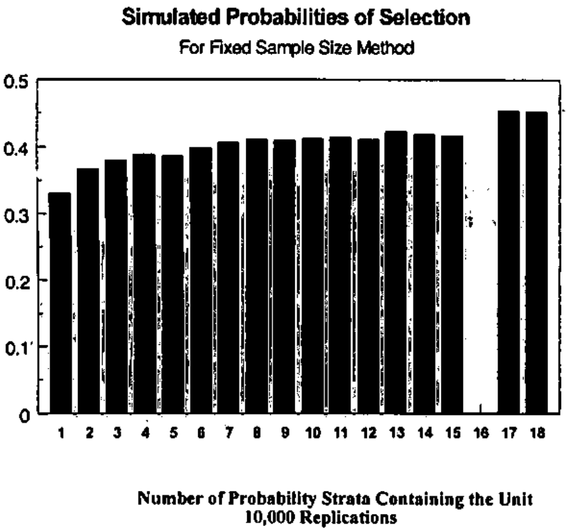


Figure 2. Simulated probabilities of selection for the fixed-sample-size method-California

6. CALIBRATION

The problem with both t_M and t_P (or t_{HT}) is that they are often not very good estimators for T in term of precision (variance). One of the properties of single-frame, stratified simple random sampling is that the conventional expansion estimator estimates the stratum population size perfectly (*i.e.*, with zero variance). In our multiple frame set up, however, neither t_M nor t_P will estimate the N_{fh} perfectly in most applications.

Let us define $w_i^0 = n_{(i)}/E[n_{(i)}]$ as the *original sampling weight* of unit i in t_M. Similarly, $w_i^0 = 1/\max_f\{p_{if}\}$ in t_P and $1/\pi_i$ more generally for a Horvitz-Thompson estimator. Bankier (1986) proposed raking to create a set of adjusted weights such that

$$\sum_{i \in S_{fh}} w_i^C = N_{fh} \tag{2}$$

for each stratum h in every frame f, where S_{fh} is that part of the sample that is in stratum h of frame f regardless of the frame(s) from which the units were selected.

Deville and Särndal (1992) call (2) a *calibration equation*. They point out that there are a number of ways to compute the *calibration weights*, the w_i^C, so that equation (2) is satisfied and w_i^C/w_i^0 is in some sense close to 1 for all i. One method is raking as suggested by Bankier (1986). Another method, discussed at length by Deville and Särndal (1992), uses least squares. Either way, the resulting estimator

$$t_C = \sum_{i \in S} w_i^C y_i,$$

where S denotes the entire sample, will be nearly design unbiased because w_i^C/w_i^0 is close to 1 for all i.

The estimator t_C is also unbiased under the model:

$$y_i = \beta_0 + \sum_{f=1}^{F} \sum_{h=2}^{H_f} d_{ifh}\,\beta_{fh} + \epsilon_i, \tag{3}$$

where the dummy variable, d_{ifh}, is 1 when unit i is in stratum h of frame f (sampled or not) and zero otherwise, while ϵ_i is a random variable with a mean of zero. The β_0 and the β_{fh} are unknown constants (β_0 represents the mean y-value for a unit in the first stratum of every frame; that is why the second sum excludes $h = 1$). The same d_{ifh} values apply to every survey item (y) of interest, while the β values change with the survey item. For many survey items, β_{fh} values will be zero when frame f (say, grain stocks) is irrelevant to the item (say, planted oat acres).

Isaki and Fuller (1982) call the model expectation of the design mean squared error of t_C the "anticipated mean squared error" of the estimator. This value is of most use at the planning stage of a sample survey.

If the model in equation (3) holds, and the ϵ_i are uncorrelated, then the anticipated mean squared error of t_C is

$$\begin{aligned}
E_\epsilon[\text{MSE}_D(t_C)] &= E_\epsilon\{E_D[\textstyle\sum_s w_i^C y_i - \sum_P y_i)^2]\} \\
&= E_D\{E_\epsilon[(\textstyle\sum_s w_i^C y_i - \sum_P y_i)^2]\} \\
&= E_D\{E_\epsilon[(\textstyle\sum_s w_i^C \epsilon_i - \sum_P \epsilon_i)^2]\} \\
&= E_D\{\textstyle\sum_s [(w_i^C)^2 - 2w_i^C] E_\epsilon(\epsilon_i^2)\} + \sum_P E_\epsilon(\epsilon_i^2) \\
&\approx E_D\{\textstyle\sum_s [(1/\pi_i)^2 - 2/\pi_i] E_\epsilon(\epsilon_i^2)\} + \sum_P E_\epsilon(\epsilon_i^2) \\
&= \textstyle\sum_P (1/\pi_i - 1) E_\epsilon(\epsilon_i^2),
\end{aligned} \tag{4}$$

since $w_i^C \approx 1/\pi_i$. It is of some interest to note that using Poisson, collocated, and systematic PPS sampling result in estimators with approximately equal anticipated mean squared errors asymptotically. This surprising result is in part due to the nature of a calibrated estimator, but it is also a repercussion of the fact that when we take the design expectation of the approximate model variance in the last line of equation (4), we average over all possible samples and remove the biggest source of variation among the three sampling designs.

Now suppose we had used stratified simple random sampling and selected unit i with probability $p_{if} \le \pi_i$, where f is the frame relevant to y. It is not hard to show that the anticipated variance of the simple expansion estimator would have been $\sum_P (1/p_{if} - 1) E_\epsilon(\epsilon_i^2)$, which is at least as large as the right hand side of equation (4). Thus, there are gains – in large samples, at least – from "integrating" the samples from various frames as we have effectively done. How large the samples must be in practice for the asymptotic results to be relevant is unclear. At the very least, the sample size must be many times the number of model parameters in equation (3).

A few words on mean squared error estimation for t_C are in order. The mean squared error estimator advocated by Deville and Särndal (1992) – an estimator with both good design and model-based properties – can not be implemented unless the joint selection probability (π_{ij}) for every pair of sample units (i and j) is known. Among the designs we have discussed, these probabilities are easily calculated only for the Poisson variant of PRN (where $\pi_{ij} = \pi_i\pi_j$).

As we have observed in equation (4), the anticipated mean squared error of the calibration estimator is the same under Poisson PRN, collocated PRN, and systematic PPS sampling. This suggests that the Poisson mean squared error estimator may be reasonable under each of the three designs. A stronger model-driven argument exists for this contention, but will not be made here.

7. DISCUSSION

In the last section, it was pointed out that if calibration weights were designed to satisfy equation (2), the resulting estimator would be unbiased under the model in equation (3). In many applications, there may be a more appropriate model on which to base calibration than the one in equation (3). For example, if there was a continuous control variable used to stratify a particular frame, it makes more sense to use that variable directly in the model rather than indirectly through frame/stratum identifiers.

Raking is a form of calibration under a particular model. With that in mind, it makes sense to use the most reasonable model available. Least squares has the advantage over raking that it can easily be applied to continuous control variables. Singh and Mohl (1996) provide an extensive review of alternative calibration algorithms including an extension of raking to continuous variables. An intriguing least-squares variant missed by Singh and Mohl (1996) can be found in Brewer (1994).

Many economic and agricultural surveys employ rotating sample designs. This has proved an effective way to balance cost and burden considerations. Although our empirical findings demonstrated an advantage of the systematic PPS methodology in terms of meeting target sample sizes, the three PRN designs are much more conducive to sample rotation. See, for example, Ohlsson (1995) on this topic. Moreover, with the PRN methods, one can integrate different frames at different times of the year (with systematic PPS there is no easy way to allocate the sample back to the frame of origin). This is a particularly useful property for agricultural surveys because different crops have different growing seasons.

In summary, the fixed-sample-size PRN sample design is excellent for meeting target sample sizes but is hard to use in practice because selection probabilities are usually unknown and must be simulated. The systematic PPS design is very good at meeting target sample sizes but is difficult to incorporate into a sample rotation scheme. Moreover, mean squared error estimation requires invocation of model assumptions. Our empirical example shows that collocated sampling may only be slightly better than Poisson at meeting target sample sizes. It should be recognized, however, that other configurations of the frames, strata, and sampling fractions may produce different results. Moreover, collocated sampling is conducive to rotation schemes, like Poisson sampling. On the other hand, like PPS sampling, it requires the assumption of a model to estimate mean squared error.

Finally, setting p_{if} or n_{if} targets is a popular, but indirect, means of controlling the variance of the estimator t_C associated with each frame. These targets lead to our *ad hoc* decision to set π_i equal to $\max_f\{p_{if}\}$. A more direct strategy would be to set (asymptotic) anticipated variance targets for each frame estimator using equation (4) and postulated values for the $E_\epsilon(\epsilon_i^2)$. One could then choose, say, the set of π_i that minimizes the expected sample size yet satisfy these variance targets. A similar approach is taken by Amrhein, Fleming, and Bailey (1997) who use Chromy's algorithm in a manner analogous to Sigman and Monsour (1995). Poisson PRN, collocated PRN, and systematic PPS sampling remain three viable alternatives for selecting the sample once optimal π_i are determined.

REFERENCES

AMRHEIN, J.F., FLEMING, C.M., and BAILEY, J.T. (1997). Determining the probabilities of selection in a multivariate probability proportional to size sample design. In *Proceedings: Symposium 97: New Directions in Surveys and Censuses.* Statistics Canada. To appear.

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

BREWER, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics,* 10(1)A, 213-233.

DEVILLE, J-C., and SÄRNDAL, C.-E. (1992). Calibration estimator in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association,* 77, 89-96.

KOTT, P.S., and FETTER, M.J. (1997). A multi-phase sample design to co-ordinate surveys and limit response burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association.* To appear.

LAVALLÉE, P., and HIDIROGLOU, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.

OHLSSON, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: Wiley, 153-169.

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

SIGMAN, R.S., and MONSOUR, N.J. (1995). Selecting samples from list frames of businesses. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: Wiley, 133-152.

SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

SKINNER, C.J., HOLMES, D.J., and HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 3, 333-347.

SWEET, E., and SIGMAN, R.S. (1995). *User Guide for the Generalized* SAS *Univariate Stratification Program*, Economical Statistical Methods and Programming Division, Bureau of the Census, U.S. Department of Commerce, Report number ESM-9504.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 11-20
Statistics Canada

Use of Auxiliary Information for Two-phase Sampling

M.A. HIDIROGLOU and C.-E. SÄRNDAL[1]

ABSTRACT

Two-phase sampling designs offer a variety of possibilities for use of auxiliary information. We begin by reviewing the different forms that auxiliary information may take in two-phase surveys. We then set up the procedure by which this information is transformed into calibrated weights, which we use to construct efficient estimators of a population total. The calibration is done in two steps: (i) at the population level; (ii) at the level of the first-phase sample. We go on to show that the resulting calibration estimators are also derivable via regression fitting in two steps. We examine these estimators for a special case of interest, namely, when auxiliary information is available for population subgroups called calibration groups. Poststrata are the simplest example of such groups. Estimation for domains of interest and variance estimation are also discussed. These results are illustrated by applying them to two important two-phase designs at Statistics Canada. The general theory for using auxiliary information in two-phase sampling is being incorporated into Statistics Canada's Generalized Estimation System.

KEY WORDS: Generalized regression; Two-phase sampling; Model assisted approach; Domain estimation; Calibration factors.

1. INTRODUCTION

Two-phase sampling is a powerful and cost-effective technique. It was first proposed by Neyman (1938). In Cochran's (1977) book, and in its two earlier editions dated 1953 and 1963, one finds basic results for two-phase sampling, including the simplest regression estimators for such designs. This paper takes a broader outlook and proposes a general approach to the use of auxiliary information in two-phase survey designs. Our main references are Särndal and Swensson (1987), Särndal, Swensson and Wretman (1992) and Dupont (1995). Recent related work includes Breidt and Fuller (1993), who presented computationally efficient estimation procedures for three-phase sampling in the presence of auxiliary information. Chaudhuri and Roy (1994) studied optimality properties of the well-known simpler regression estimators for two-phase sampling. Binder (1996) described a simple linearization procedure to estimate variances of nonlinear estimators. His procedure can be applied to any sampling design, including two-phase-sampling. Throughout this paper, we assume *arbitrary* sampling designs for each of the two phases.

Single-phase sampling involves the use of one layer of information for estimation. In two-phase sampling, however, one has to consider two layers of information. This complicates matters, and it is not clear-cut how best to exploit the combined information from the two sources. Two approaches are considered in this paper for building estimators based on auxiliary information. These are the *calibration approach* and the *generalized regression approach*. We show that the generalized regression approach can be viewed as a special case of the calibration approach. The two approaches are examined under a common structure for the auxiliary information. It assumes that information exists about an auxiliary vector x_1 for the units of the entire population, and about a second auxiliary vector x_2 for the units of the first phase sample. Consequently, at the level of the first phase sample, there is information about both vectors, x_1 and x_2.

The *generalized regression approach,* as applied to two-phase sampling, is discussed in Särndal *et al.* (1992). These authors develop the general regression estimator for two-phase sampling, assuming arbitrary sampling designs in each of the two phases. Two regression fits are carried out. A "bottom level" regression is fitted to produce predicted values up to the level of the first phase sample, using the auxiliary information available for this step. Next, a "top level" regression is fitted to produce predicted values up to the entire population level, using the information appropriate for this step. The two sets of predicted values are used to build a generalized regression estimator.

The *calibration approach* focuses on the weights given to the units for purposes of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. In two-phase sampling the two levels of information imply two consecutive calibrations. The first phase of calibration uses the auxiliary information available (at least population counts) at the level of the entire

[1] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; and C.-E. Särndal, University of Montreal, and Statistics Canada.

for $k \varepsilon s_2$, and

$$T_2 = \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}'_k}{C_{2k}} \qquad (3.12)$$

Again, some g_k^* may be zero or negative, but always positive g_k^* can be ascertained by adding to (3.8) the inequality constraints $w_k^* > 0$ for $k \varepsilon s_2$.

Having determined the overall weights $\tilde{w}_k^*$ by equation (3.9), the estimator of Y is given by

$$\hat{Y} = \sum_{s_2} \tilde{w}_k^* y_k \qquad (3.13)$$

Remark 3.1 A potential problem with the above approach is that some of the g_{1k}'s may be negative or even zero. If this occurs, (3.7) is not a proper distance measure. Some of the important applications, such as poststatification, do not have this problem as their associated g_{1k}'s are always greater than zero. If all the g_{1k}'s are greater than zero, then the minimization criterion given by (3.7) is acceptable. Otherwise, we have to modify it. One possible modification is to impose on the above-mentioned constraints that the w_{1k}'s are positive for $k \varepsilon s_1$. Another possible modification is to replace C_{2k} in (3.7) by

$$C_{2k}^* = C_{2k} \frac{\tilde{w}_{1k}}{w_{1k}}.$$

Then

$$\frac{C_{2k}^*}{\tilde{w}_{1k} w_{2k}} = \frac{C_{2k}}{w_k^*},$$

which is always positive. The resulting g_k^*-factors in (3.9) can be shown to be $g_k^* = g_{1k} + g_{2k} - 1$, where g_{1k} is given as before by (3.5), and g_{2k} by (3.11) provided that we instead define T_2 as

$$T_2 = \sum_{s_2} \frac{w_k^* \boldsymbol{x}_k \boldsymbol{x}'_k}{C_{2k}}.$$

It is our opinion that in most applications the choice between the multiplicative $g_k^* = g_{1k} g_{2k}$ and the additive form $g_k^* = g_{1k} + g_{2k} - 1$ would have little effect on the resulting estimates. That is, we believe the two point estimates would be very close, and so would be their associated estimates of variance.

Remark 3.2: Bounding the weights ordinarily has negligible impact on the estimates. Recent experience with calibration for single phase designs, Stukel, Hidiroglou, and Särndal (1996), has shown that mildly different sets of g-weights lead to point estimates that differ very little. Some recently developed computer software for calibration, for example, the software described in Deville *et al.* (1993), minimizes a distance function such that the resulting g-factors are guaranteed to be bounded from above and from below.

Remark 3.3: The auxiliary data in Table 1 can be used in several ways for two-phase calibration. Considering in particular the second-phase calibration equation defined by (3.8), three different specifications of the vector $\boldsymbol{x}_k$ are: (i) $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$; (ii) $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$; and (iii) $\boldsymbol{x}_k = \boldsymbol{x}_{1k}$. We comment on these possibilities, assuming for each of these that a first-phase calibration has been carried out, resulting in the first-phase calibrated weights (3.4).

The case (i) specification $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$, recommended in Särndal *et al.* (1992), capitalizes on all the available information. Thus, in this respect case (i) is ideal. Cases (ii) and (iii) disregard some available information. Case (ii) is sometimes of interest, despite some loss of information; an example is given in Section 7.1. Case (iii) implies that the data $\{\boldsymbol{x}_{2k} : k \varepsilon s_1\}$ are observed, but not used: we do not further consider this case. We call $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$ the *full vector* and $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$ the *reduced vector.*

Second-phase calibration on the reduced vector $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$ can be carried out without significant loss of information if $\boldsymbol{x}_{2k}$ is a good *substitute* for $\boldsymbol{x}_{1k}$, as also observed by Dupont (1995). However, if $\boldsymbol{x}_{1k}$ *complements* $\boldsymbol{x}_{2k}$, then the full vector $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$ should clearly be used in the calibration defined by (3.7). Otherwise, significant loss of information and increased variance may result.

Remark 3.4: Both the full and the reduced $\boldsymbol{x}_k$-vectors lead to overall weights $\tilde{w}_k^*$ calibrated on $\boldsymbol{x}_{2k}$ from s_2 to s_1. This means that $\sum_{s_2} \tilde{w}_k^* \boldsymbol{x}_{2k} = \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{2k}$, because (3.8) holds, and $\boldsymbol{x}_{2k}$ is contained in $\boldsymbol{x}_k$. However, there exists a difference between the full and reduced vector specifications with respect to the calibration on $\boldsymbol{x}_{1k}$. If the full vector specification is used in phase two, the resulting overall weights $\tilde{w}_k^*$ are calibrated on $\boldsymbol{x}_{1k}$ from s_2 to s_1, and from s_1 to U. This means that $\sum_{s_2} \tilde{w}_k^* \boldsymbol{x}_{1k} = \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k}$. In contrast, if the reduced vector specification is used, the resulting overall weights $\tilde{w}_k^*$ **are** calibrated on $\boldsymbol{x}_{1k}$ from s_1 to U by virtue of the first-phase calibration. That is $\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k}$. However, they are **not** calibrated from s_2 to s_1, because $\boldsymbol{x}_{1k}$ is not present in the second-phase calibration. Hence, $\sum_{s_2} \tilde{w}_k^* \boldsymbol{x}_{1k} \neq \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k}$. Thus if the survey requires a weight system that will reproduce the known $\sum_U \boldsymbol{x}_{1k}$, then the full vector specification must be used.

So far, we have focused on the general framework for calibration with two levels of auxiliary information. This framework does not reveal the many interesting forms that the estimator $\hat{Y}$ given by (3.13) may take for specific cases of auxiliary information. Some illustrations are given in Section 7. We first address three issues that are of practical interest in virtually every major survey: (i) poststratification or, more generally, the presence of auxiliary information for population subgroups (Section 5), (ii) estimation for domains of interest (Section 6), and (iii) the construction of variance estimates (Section 6).

4. THE TWO-PHASE CALIBRATION ESTIMATOR VIEWED AS A REGRESSION ESTIMATOR

An alternative expression for the calibration estimator (3.13) is given by formula (4.1) below. This expression links it exactly with the regression estimator for two-phase designs introduced in Särndal *et al.* (1992, chapter 9).

Theorem 4.1: When the overall calibrated weights $\tilde{w}_k^*$ are determined by (3.9), the calibration estimator (3.13) is identical to the two-phase regression estimator given by

$$\hat{Y} = \sum_U \hat{y}_{1k} + \sum_{s_1} w_{1k}(\hat{y}_{2k} - \hat{y}_{1k}) + \sum_{s_2} w_k^*(y_k - \hat{y}_{2k}) \quad (4.1)$$

where $\hat{y}_{1k}$ and $\hat{y}_{2k}$ are successive regression predictions such that

$$\hat{y}_{1k} = x_{1k}'\hat{B}_1 \quad (4.2)$$

with

$$\hat{B}_1 = T_1^{-1}\left\{\sum_{s_1} \frac{w_{1k}x_{1k}\hat{y}_{2k}}{c_{1k}} + \sum_{s_2} \frac{w_k^* x_{1k}(y_k - \hat{y}_{2k})}{c_{1k}}\right\} \quad (4.3)$$

where T_1 is given by (3.6), and

$$\hat{y}_{2k} = x_k'\,\hat{B}_2 \quad (4.4)$$

with

$$\hat{B}_2 = T_2^{-1}\sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} x_k y_k}{c_{2k}} \quad (4.5)$$

where T_2 is given by (3.12).

The proof for Theorem 4.1 uses some tedious but straightforward algebra and is not presented here.

We now show that (4.1) can be constructed via regression estimation in two steps. For the first step, suppose that the variable of interest y_k were observed for the full first-phase sample s_1. The auxiliary information on x_{1k} is available for $k \in s_1$ and the population total $\sum_U x_{1k}$ is known. The resulting regression estimator of $Y = \sum_U y_k$ would then be given by

$$\begin{aligned}\hat{Y} &= \sum_U \hat{y}_{1k}^0 + \sum_{s_1} w_{1k}\left(y_k - \hat{y}_{1k}^0\right)\\ &= \sum_{s_1} w_{1k}\, y_k + \left(\sum_U \hat{y}_{1k}^0 - \sum_{s_1} w_{1k}\,\hat{y}_{1k}^0\right) \quad (4.6)\end{aligned}$$

In the last expression, the first term represents the (hypothetical) first-phase Horvitz-Thompson estimator of Y. The second and third terms represent a regression adjustment, where $\hat{y}_{1k}^0$ is the predictor of y_k based on the fitted regression of y_k on x_{1k} for $k \in s_1$. That is, $\hat{y}_{1k}^0 = x_{1k}'\hat{B}_1^0$, with

$$\hat{B}_1^0 = T_1^{-1}\sum_{s_1} \frac{w_{1k}\, x_{1k}\, y_k}{c_{1k}}.$$

Note that $\sum_U \hat{y}_{1k}^0 = (\sum_U x_{1k})'\,\hat{B}_1^0$ where $\sum_U x_{1k}$ is known. However, none of the terms in (4.6) can be computed directly, because y_k is only observed for the second-phase sample. A second step of regression estimation is thus necessary. It is carried out by replacing the unknown $\sum_{s_1} w_{1k}y_k$ in (4.6) by its conditional regression estimator

$$\sum_{s_1} w_{1k}\hat{y}_{2k} + \sum_{s_2} w_k^*\,(\,y_k - \hat{y}_{2k}\,) \quad (4.7)$$

where $\hat{y}_{2k} = x_k'\hat{B}_2$, with $\hat{B}_2$ given by (4.5), is the predictor of y_k based on the regression of y_k on x_k, known up to s_1. Next, the vector $\hat{B}_1^0$ required for computing $\hat{y}_{1k}^0$ contains a known matrix T_1 and an unknown vector

$$\sum_{s_1} \frac{w_{1k}\, x_{1k} y_k}{c_{1k}}.$$

Using a regression estimator for this unknown vector, we obtain $\hat{B}_1$ given by (4.3) as a replacement for $\hat{B}_1^0$. These two substitutions in (4.6) lead to the two-phase regression estimator given by (4.1), which is identical to the calibration estimator (3.13).

Remark 4.1: A more direct alternative to $\hat{B}_1$ in (4.3) would be to use only the second-phase sample. This would have produced

$$\hat{B}_{1,\text{alt}} = \left(\sum_{s_2} \frac{w_k^*\, x_{1k}x_{1k}'}{c_{2k}}\right)^{-1} \sum_{s_2} \frac{w_k^*\ x_{1k}y_k}{c_{2k}}$$

The resulting predictions $\hat{y}_{1k,\text{alt}} = x_{1k}'\hat{B}_{1,\text{alt}}$ would be replacing $\hat{y}_{1k}$ in (4.1). However, the resulting regression estimator is not identical to (3.13) and is a less efficient alternative, because $\hat{B}_{1,\text{alt}}$ uses less x_{1k}-information than $\hat{B}_1$.

5. CALIBRATION GROUPS

In this Section we apply the results of Sections 3 and 4 to the important case where the auxiliary data in Table 1 include information about mutually exclusive and exhaustive subsets of the population U, and of the first-phase sample s_1. The population subsets are denoted by $U_i, i = 1, \ldots, I,$ and the first-phase subsets by $s_{1j}, j = 1, \ldots, J$. Such subsets are called calibration groups, for reasons that will become clear later in this Section. Simple examples of calibration groups are poststrata.

Two vectors denoted Δ_{1k} and Δ_{2k} will be used to specify the membership of a given unit k in the calibration groups U_i and s_{1j}, respectively. These group identifiers are

$$\Delta_{1k} = (\delta_{11k}, \ldots, \delta_{1ik}, \ldots, \delta_{1Ik})' \quad (5.1)$$

with

$$\delta_{1ik} = \begin{cases} 1 & \text{if } k \varepsilon U_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, ..., I \tag{5.2}$$

and

$$\Delta_{2k} = (\delta_{21k}, ..., \delta_{2jk}, ..., \delta_{2Jk})' \tag{5.3}$$

with

$$\delta_{2jk} = \begin{cases} 1 & \text{if } k \varepsilon s_{1j} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, ..., J \tag{5.4}$$

Besides the group membership information, which is qualitative and specified by Δ_{1k} and Δ_{2k}, there may exist information for the unit k about quantitative (continuous or discrete) variables. We call them *supplementary auxiliary variables*. For example, categorical information about a unit (enterprise) in a business survey may consist of an industry code or a geographical location code. In addition, quantitative variable information may also be available concerning the number of employees or the gross business income of the unit. Some of these supplementary auxiliary variables may be known up to the level of the population, and others up to the level of the first-phase sample.

We assume in this Section that the vector $\boldsymbol{x}_{1k}$, used in calculating the first-phase g-factors, has the structure

$$\boldsymbol{x}'_{1k} = \Delta'_{1k} \otimes z'_{1k} \tag{5.5}$$

where z_{1k} of dimension Q_1 is the vector of supplementary auxiliary variables available for the first-phase sample. The information requirements in Table 1 apply to the vector $\boldsymbol{x}_{1k}$. This implies that we must know either the group membership specified by Δ_{1k} and the value of z_{1k} for every $k \in U$, or the total $\sum_{U_i} z_{1k}$ separately for each group, $i = 1, ..., I$.

When $\boldsymbol{x}_{1k}$ has the form given by (5.5), the first-phase g-factors g_{1k} in (3.5) can be obtained by a group by group calculation. The $\boldsymbol{T}_1$ matrix to be inverted, given by (3.6), is block diagonal and of dimension IQ_1 by IQ_1. The typical diagonal block, denoted as $\boldsymbol{T}_{1i}$ of dimension Q_1 by Q_1, is given by

$$\boldsymbol{T}_{1i} = \sum_{s_{1i}} \frac{w_{1k}\, z_{1k} z'_{1k}}{c_{1k}} \tag{5.6}$$

for $i = 1, ..., I$. The resulting inverse of $\boldsymbol{T}_1$ is also block diagonal with diagonal matrices $\boldsymbol{T}_{1i}^{-1}$. The off diagonal blocks of the inverse of $\boldsymbol{T}_1$ are zero matrices. So we obtain from (3.6)

$$g_{1k} = 1 + \left(\sum_{U_i} z_{1k} - \sum_{s_{1i}} w_{1k} z_{1k}\right)' \boldsymbol{T}_{1i}^{-1} \frac{z_{1k}}{c_{1k}} \tag{5.7}$$

for $k \in s_{1i}$, $i = 1, ..., I$, where $\boldsymbol{T}_{1i}$ is given by (5.6). Note that the resulting weights $\tilde{w}_{1k}$ are the same as those obtained by carrying out the first-phase calibration group by group, calibrating for group i on the known total $\sum_{U_i} z_{1k}$. That is, $\sum_{s_{1i}} \tilde{w}_{1k} z_{1k} = \sum_{U_i} z_{1k}$ for $i = 1, ..., I$. It is thus fitting to call the groups U_i *first-phase calibration groups*.

Now consider the second-phase g-factors g_{2k} given by (3.11). They are based on the auxiliary vectors $\boldsymbol{x}_k$, required to be known for the units $k \in s_1$. We assume that $\boldsymbol{x}_k$ contains information about the second-phase groups so that

$$\boldsymbol{x}'_k = \Delta'_{2k} \otimes z'_k \tag{5.8}$$

where Δ_{2k} is the second-phase group identifier, and z_k is the value of a vector of supplementary auxiliary variables available for $k \in s_1$. Since the requirements in Table 1 apply, it follows that Δ_{2k} (the second-phase group membership) and the value of z_k (the supplementary auxiliary vector) must be known for every $k \in s_1$. Here z_k may contain some or all of the information in $\boldsymbol{x}_{1k}$ given by (5.5), and any other information available for the units $k \in s_1$.

When $\boldsymbol{x}_k$ has the structure (5.8), the factors g_{2k} can also be obtained through a group by group calculation. This simplification is a result of the fact that the matrix to be inverted in (3.11) is block diagonal. We obtain

$$g_{2k} = 1 + \left(\sum_{s_{1j}} \tilde{w}_{1k} z_k - \sum_{s_{2j}} \tilde{w}_{1k} w_{2k} z_k\right)' \boldsymbol{T}_{2j}^{-1} \frac{z_k}{c_{2k}} \tag{5.9}$$

for $k \in s_{2j} = s_2 \cap s_{1j}$, $j = 1, ..., J$, where

$$\boldsymbol{T}_{2j} = \sum_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} z_k z'_k}{c_{2k}} \tag{5.10}$$

The resulting overall weights $\tilde{w}_k^* = w_k^* g_k^*$ where $g_k^* = g_{1k} g_{2k}$ are the same as those obtained by carrying out the second-phase calibration group by group, calibrating for group j on the known quantity $\sum_{s_{1j}} \tilde{w}_{1k} z_k$. That is, $\sum_{s_{2j}} \tilde{w}_k^* z_k = \sum_{s_{1j}} \tilde{w}_{1k} z_k$ for $j = 1, ..., J$. The groups s_{1j} are called *second-phase calibration groups*. We now have a procedure for computing g_{1k} and g_{2k} group by group using (5.7) and (5.9). The total Y is still estimated according to (3.13).

6. DOMAIN ESTIMATION AND VARIANCE ESTIMATION

The preceding sections dealt with estimation of the total of y at the entire population level. In most surveys, there is also a need to provide estimates for various subpopulations or domains of interest. Requests for domain estimates can be made either before or after the sampling stage of the survey. Auxiliary information is essential for domains. A precise domain estimate may be obtained (even for small domains) if: (i) calibration groups and domains of interest agree closely, and (ii) the auxiliary variables exhibit a strong regression relationship with the variable(s) of interest.

Denote by $U_d \, (U_d \subseteq U)$ any domain of the population U for which an estimate is required. The y-total for the domain U_d is defined by $Y(d) = \sum_{U_d} y_k = \sum_U y_k(d)$ with $y_k(d) = y_k$ if $k \varepsilon U_d$ and $y_k(d) = 0$ if $k \notin U_d$.

The estimator of $Y(d)$ is

$$\hat{Y}(d) = \sum_{s_2} \tilde{w}_k^* \, y_k(d) \qquad (6.1)$$

where the overall calibrated weights $\tilde{w}_k^* = w_k^* \, g_k^*$ may be calculated group by group as described in Section 5. The calibration factors g_{1k} and g_{2k} are calculated using all relevant available auxiliary information, specified as in Table 1. So in this sense, the resulting overall calibrated weights $\tilde{w}_k^*$ are the best possible ones. Note that these weights are independent of the particular domains requiring estimation in the survey.

The estimator of the variance for the domain total estimator $\hat{Y}(d)$ is obtained using a design-based approach. This means that the variance is interpreted with reference to repeated draws of samples s_1 and s_2. Details for the derivation of this variance are given in Särndal *et al.* (1992) (Result 9.7.1, p. 362). The first order and second order inclusion probabilities enter into the weights used in the variance formula. The weights associated with the first-phase sample are $w_{1k} = 1/\pi_{1k}$ and $w_{1k\ell} = 1/\pi_{1k\ell}$ with $\pi_{1k\ell} = P(k \text{ and } \ell \in s_1)$. The weights $w_{2k} = 1/\pi_{2k}$ and $w_{2k\ell} = 1/\pi_{2k\ell}$ with $\pi_{2k\ell} = P(k \text{ and } \ell \in s_2 \,|\, s_1)$ denote their second phase counterparts. Two sets of regression residuals, one for each phase, are also required. The estimator of the variance of $\hat{Y}(d)$ is given by

$$v\{\hat{Y}(d)\} =$$

$$\sum_{k \varepsilon s_2} \sum_{\ell \varepsilon s_2} w_{2k\ell}(w_{1k} w_{1\ell} - w_{1k\ell})(g_{1k} e_{1k}(d))\,(g_{1\ell} e_{1\ell}(d)) + \qquad (6.2)$$
$$\sum_{k \varepsilon s_2} \sum_{\ell \varepsilon s_2} w_{1k} w_{1\ell}(w_{2k} w_{2\ell} - w_{2k\ell})\,(g_{2k} e_{2k}(d))(g_{2\ell} e_{2\ell}(d))$$

Note that for $k = \ell$ we have $w_{1k\ell} = w_{1k}$, and $w_{2k\ell} = w_{2k}$ in (6.2). We now specify the regression residuals in (6.2) assuming that there are first-phase calibration groups $U_i, i = 1, ..., I,$ and second-phase calibration groups $s_{1j}, j = 1, ..., J,$ as explained in Section 5. We denote the associated sample subsets as follows: $s_{2i} = s_2 \cap U_i$; $s_{2j} = s_2 \cap s_{1j}$. The required residuals in (6.2) are, for $k \in (s_{2i} \cap U_d)$,

$$e_{1k}(d) = y_k(d) - z_{1k}' \hat{\boldsymbol{B}}_{1i}(d) \qquad (6.3)$$

and, for $k \in (s_{2j} \cap U_d)$

$$e_{2k}(d) = y_k(d) - z'_k \hat{\boldsymbol{B}}_{2j}(d) \qquad (6.4)$$

The estimated regression vectors $\hat{\boldsymbol{B}}_{1i}(d)$ and $\hat{\boldsymbol{B}}_{2j}(d)$ are

$$\hat{\boldsymbol{B}}_{1i}(d) = \boldsymbol{T}_{1i}^{-1}$$

$$\left\{ \sum_{s_{1i}} \frac{w_{1k} z_{1k} \hat{y}_{2k}(d)}{C_{1k}} + \sum_{s_{2i}} \frac{w_k^* \; z_{1k}(y_k(d) - \hat{y}_{2k}(d))}{C_{1k}} \right\} (6.5)$$

where $\boldsymbol{T}_{1i}$ is given by (5.6), and

$$\hat{\boldsymbol{B}}_{2j}(d) = \boldsymbol{T}_{2j}^{-1} \sum_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} z_k y_k(d)}{C_{2k}} \qquad (6.6)$$

with $\boldsymbol{T}_{2j}$ given by (5.10), and

$$\hat{y}_{2k}(d) = z_k' \hat{\boldsymbol{B}}_{2j}(d) \text{ for } k \in (s_{ij} \cap U_d).$$

Remark 6.1: Note that for each new domain of interest, the variance estimator (6.2) requires two new sets of domain dependent residuals, $e_{1k}(d)$ and $e_{2k}(d)$. Moreover, these are required for *all* of the units k in the second-phase sample s_2, including units outside the domain. Variance estimation for domains can therefore be cumbersome.

Remark 6.2: In practice the computation of estimated variances is seldom carried out as a double sum. For some important designs, the double sums reduce, after some algebraic manipulation, to single sum expressions. Examples of this occur for single sampling and for stratified single random sampling in both phases. Explicit algebraic developments for the variances have been given the former case by Särndal *et al.* (1992), and in the later case by Hidiroglou (1995), and Binder, Babyak, Brodeur, Hidiroglou and Jocelyn (1997).

7. APPLICATIONS WITH POSTSTRATIFICATION AT THE FIRST PHASE

7.1 The Case of the Tax Sample at Statistics Canada

An application of the calibration group approach in section 5 has been in use at Statistics Canada, in the two-phase design for sampling of tax records. The example is important because it provides the extension to two-phase designs of the traditional postratification technique as used in a single phase design. The sampling procedure, the post-stratification criteria, and the estimators are described in Armstrong and St-Jean (1994). We now show how these estimators are obtained as special case of the technique in section 5. The sampling design, in each phase, is stratified Bernouilli, carried out with the permanent random number technique. The two stratifications are based on different criteria. The realized sample sizes are random at each phase on account of the Bernouilli sampling. To offset the resulting tendency toward an increased variance, poststratification is carried out at both phases of sampling. The two poststratification criteria are different. We have in effect two crossing poststratifications. In the terminology of section 5, the first phase poststrata are the first-phase calibration groups. They are denoted as U_i; $i = 1, ..., I$, and the group membership of a unit k is indicated by the vector by Δ_{1k} given by (5.1). The second phase poststrata are the second phase calibration groups. They are denoted as $s_{1j}, j = 1, ..., J$ and the corresponding membership of a unit k is indicated by the vector Δ_{2k} given by (5.3).

The first-phase calibration is carried out using the information about the first-phase poststrata sizes, N_i. In this survey design, there is no supplementary information, so $z_{1k} = 1$ for all k in (5.5), yielding $x_{1k} = \Delta_{1k}$. Specifying $C_{1k} = 1$ for all k we obtain from (5.7) that

$$g_{1k} = N_i / \hat{N}_{1i} \qquad (7.1)$$

for all $k \in s_{1i}$ where $\hat{N}_{1i} = \sum_{s_{1i}} w_{1k}$ estimates the known first-phase poststratum count N_i, and $s_{1i} = s_i \cap U_i$ denotes the part of the first-phase sample s_1 that falls in the first-phase poststratum U_i.

We arrive at the estimator of Armstrong and St-Jean (1994) by carrying out the second-phase calibration with $x_k = \Delta_{2k}$, that is, we have $z_k = 1$ for all k in (5.8). This is a reduced x_k-vector specification since it does not involve x_{1k}. Specifying $C_{2k} = 1$ for all $k \in s_{1i}$, and using (5.9) and (3.10), we obtain the overall calibrated weights

$$g_k^* \doteq \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1j}}{\hat{N}_{2j}} \qquad (7.2)$$

for all $k \varepsilon s_{2ij}$, where

$$\hat{N}_{1j} = \sum_{i=1}^{I} \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{1ij}; \hat{N}_{2j} = \sum_{i=1}^{I} \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{2ij} \qquad (7.3)$$

with $\hat{N}_{1ij} = \sum_{s_{1ij}} w_{1k}$ and $\hat{N}_{2ij} = \sum_{s_{2ij}} w_k^*$. Here, $s_{2j} = s_2 \cap s_{1j}$ denotes the part of the second-phase sample s_2 that falls in the second-phase poststratum s_{1j}, and $s_{1ij} = U_i \cap s_{1j}$; $s_{2ij} = s_2 \cap U_i \cap s_{1j}$. It follows that the estimator of the total $Y(d)$ for a given domain U_d is given by $\hat{Y}(d) = \sum_{s_2} w_k^* g_k^* y_k(d)$, or equivalently as

$$\hat{Y}(d) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{ij}}{\hat{N}_{2j}} \sum_{s_{2ij}} w_k^* y_k(d).$$

The estimated variance requires two types of residuals that are easily obtained from the general expressions given in Section 6.

Alternatives exist to the reduced vector specification $x_k = \Delta_{2k}$ used for this design. We therefore examine what the estimator would look like under a full vector specification. For the first-phase calibration, as earlier, let $x_{1k} = \Delta_{1k}$ corresponding to $z_{1k} = 1$ for all k in (5.8). The first-phase g-factors g_{1k} are then given by (7.1). In this survey, information is available for assigning every unit $k \in s_1$ to one of the $I \times J$ cells formed by cross-classifying the two poststratification criteria. Therefore, the vector x_k for the second-phase calibration can be taken as

$$x_k' = \Delta_{1k} \otimes \Delta_{2k}' \qquad (7.4)$$

This is a full vector specification in that it includes the first-phase information carrier Δ_{1k}. Let us also specify $C_{2k} = 1$ for all k. Since (7.4) is of the form (5.8), the second-phase g-factors g_{2k} are obtainable group-by-group from (5.9) with $z_k = \Delta_{1k}$. The overall calibration factors are given by

$$g_k^* = \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1ij}}{\hat{N}_{2ij}} \qquad (7.5)$$

for all $k \in s_{2ij}$. Here, $\hat{N}_{1i}$ is defined in (7.1), and $\hat{N}_{1ij}$ and $\hat{N}_{2ij}$ are as in (7.3). These overall calibration factors are the product of two poststratified calibration factors. They are all positive and well defined, provided all sample cells s_{2ij} are non-empty. Collapsing of small cells s_{2ij} with relatively large non-empty cells is recommended for stable estimation. As pointed out in Remark 3.4, the overall weights obtained from (7.5) reproduce the known first-phase postrata sizes N_i, whereas those obtained from (7.2) do not.

Remark 7.1: Let us compare the calibration factors (7.2) and (7.5), resulting, respectively, from the reduced form $x_k = \Delta_{2k}$ and from the full form (7.4). Both factors are a product of two terms. The only difference lies in the second term. In both cases, the computation of the second term requires cross-classification information. That is, for every $k \in s_1$, we need to identify the cross-classification cell ij to which k belongs. In the case of the reduced vector, the cell information is pooled across the first-phase groups. For the full vector, the cell information is kept separate, and one would expect the resulting weights to be more efficient.

Remark 7.2: For the second-phase calibration, an alternative to (7.4) that also captures the information about the first-phase poststrata is to use

$$x_k' = \left(\Delta_{1k}', \Delta_{2k}' \right). \qquad (7.6)$$

Note that with this specification, there is only one calibration group in the second phase, namely the whole first-phase sample s_1.

7.2 The Case of the Canadian Survey Employment, Payrolls and Hours

The Survey on Employment Payrolls, and Hours (SEPH) covers all sectors of Canadian industry, and collects data on four principal variables: (i) salaries and payments to employees (denoted as z_2; called payrolls); (ii) number of employees (z_3; employment); (iii) hours worked by employees (y_1; hours); and (iv) summarized earnings (y_2; earnings).

SEPH (1994) uses a stratified two-phase sampling design. In the first phase, a sample of payroll deduction accounts is selected using a stratified Bernoulli sampling design with sampling rates within strata ranging from 10% to 100%. The strata are defined by region. A region is made up of one or more Canadian provinces. We describe the estimation for SEPH by considering one specific region.

For units selected in the first-phase sample, two variables are transcribed, namely, payrolls (z_2) and number of employees (z_3). In the second-phase, a simple random sample is drawn. Data on the two variables of interest, y_1 and y_2, are collected for respondents in this sample. In addition, classification by industry and province is recorded for sampled units. The first-phase sample is poststratified by employment size groups. These are used as first-phase calibration groups and denoted $U_i; i = 1, ..., I$. Their sizes denoted as N_i for $i = 1, ..., I$ are assumed known. The vector x_{1k} used for a first-phase calibration is of the form (5.5), where Δ_{1k} is given by (5.1) and $z_{1k} = 1$ for all k. We choose $C_{1k} = 1$ for all k. It follows from (5.7) that the first-phase g-factors are

$$g_{1k} = N_i / \hat{N}_{1i} \tag{7.7}$$

for all $k \in s_{1i} = s_i \cap U_i$, where $\hat{N}_{1i} = \sum_{s_{1i}} w_{1k}, i = 1, ..., I$.

We now turn to second-phase calibration. It is carried out using calibration groups $s_{1j}, j = 1, ..., J$, identified by the vector Δ_{2k} given by (5.3). These groups are based on a province by industry classification. They are constructed so that: (i) there is a strong regression relationship between y_k and the two z-variables, and that (ii) there are at least 30 observations within each group. The $J(I+2)$ dimensional $\boldsymbol{x}_k$-vector for the second-phase calibration is given by

$$\boldsymbol{x}_k' = \Delta_{2k}' \otimes (\Delta_{1k}', z_{2k}, z_{3k}) \tag{7.8}$$

This specification requires (see Table 1) that every $k \in s_1$ can be classified into one of the I by J cells formed by crossing the calibration groups in the two phases. Let $s_{2j} = s_2 \cap s_{1j}$; $s_{1ij} = s_{1j} \cap U_i$; $s_{2ij} = s_2 \cap s_{1ij}$. Also, the quantitative variable values z_{2k} (payrolls) and z_{3k} (number of employees) must be known for $k \in s_1$. The $\boldsymbol{x}_k$-vector specification given by (7.8) is full, because it incorporates $\boldsymbol{x}_{1k} = \Delta_{1k}$. A reduced vector, ignoring the first-phase groups, would be $\boldsymbol{x}_k' = \Delta_{2k}' \otimes (z_{2k}, z_{3k})$.

As in Example 7.1, we have two crossing sets of calibration groups.

Since the $\boldsymbol{x}_k$-vector (7.8) has the structure defined by (5.8), we used (5.9) to derive the second-phase g-factors for each group $j = 1, ..., J$. It follows from (7.8) that we are fitting, within each second-phase calibration group, a separate regression of y_k on $\zeta_k = (z_{2k}, z_{3k})'$ with an intercept that varies with the first-phase calibration group.

Specifying $C_{2k} = 1$ for all k, and using the additive form, $g_k^* = g_{1k} + g_{2k} - 1$, for the overall calibration factors, we obtain after some algebra

$$g_k^* = G_{1i} G_{2ij} + \boldsymbol{H}_j' \boldsymbol{T}_j^{-1} \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right)$$

for all $k \in s_{2ij}$, where

$$G_{1i} = N_i / \hat{N}_{1i}, \; G_{2ij} = \hat{N}_{1ij} / \hat{N}_{2ij}$$

$$\boldsymbol{H}_j = \sum_{i=1}^{I} \hat{N}_{1ij} G_{1i} \left(\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}} \right)$$

$$\boldsymbol{T}_j = \sum_{j=1}^{I} \sum\nolimits_{s_{2ij}} w_k^* \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right) \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right)'$$

with

$$\bar{\zeta}_{s_{1ij}} = \sum\nolimits_{s_{1ij}} \frac{w_{1k} \zeta_k}{\hat{N}_{1ij}} ; \bar{\zeta}_{s_{2ij}} = \sum\nolimits_{s_{2ij}} \frac{w_k^* \zeta_k}{\hat{N}_{2ij}} ; \hat{N}_{1ij} = \sum\nolimits_{s_{1ij}} w_{1k} ;$$

and $\hat{N}_{2ij} = \sum_{s_{2ij}} w_k^*$.

It follows that we can write the estimator (6.1) as $\hat{Y}(d) = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{Y}_{ij}(d)$ with

$$\hat{Y}_{ij}(d) = G_{1i} \hat{N}_{1ij} \{ \bar{y}_{s_{2ij}}(d) + (\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}})' \hat{\boldsymbol{B}}_j(d) \}$$

where

$$\bar{y}_{s_{2ij}}(d) = \sum\nolimits_{s_{2ij}} w_k^* y_k(d) / \hat{N}_{2ij}$$

and $\hat{\boldsymbol{B}}_j(d) = \boldsymbol{T}_j^{-1} \sum_{i=1}^{I} \sum_{s_{2ij}} w_k^* (\zeta_k - \bar{\zeta}_{s_{2ij}}) y_k(d)$.

The form of $\hat{Y}(d)$ is easy to understand. It is composed of $I \times J$ cell estimates $\hat{Y}_{ij}(d)$, each reflecting the regression of $y_k(d)$ on ζ_k. Note that the two-dimensional slope vector $\hat{\boldsymbol{B}}_j(d)$ is obtained by pooling data across the first-phase groups. This is because the specification (7.8) of $\boldsymbol{x}_k$ allows the intercept, but not the two regression slopes, to vary with the first-phase groups.

8. CONCLUSIONS

Two-phase designs have the advantage of being both economical and efficient. The present paper has provided a general theory for such designs when auxiliary information is present in each phase.

Our goal is to incorporate this two-phase survey methodology into Statistics Canada's Generalized Estimation System (GES) described in Estevao *et al.* (1995). The GES is a general purpose program that currently handles domain estimation for arbitrary single phase designs and incorporates auxiliary information in its estimation process. In this paper we have extended the basic principles of the GES, including the important idea of calibration groups, to two-phase designs.

We have illustrated the theory by showing its use in two current surveys at Statistics Canada. Given its generality, the theory has potential application to any two-phase sample design that uses auxiliary information.

REFERENCES

ARMSTRONG, J., and ST-JEAN, H. (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology*, 20, 97-106.

BINDER, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A., and JOCELYN, W. (1997). Variance Estimation for Two-phase Stratified Sampling. Contributed paper presented at the Annual Meeting of the American Statistical Association, Los Angeles.

BREIDT, J., and FULLER, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā*, 55, 297-309.

CHAUDHURI, A., and ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.

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

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DUPONT, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-136.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the Canadian Survey of Employment, Payrolls and Hours redesign. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 123-128.

HIDIROGLOU, M.A., LATOUCHE, M., ARMSTRONG, B., and GOSSEN, M. (1995). Improving survey information using administrative records: the case of the Canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.

HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 873-878.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of The American Statistical Association*, 33, 101-116.

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

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

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimation in survey sampling. *Survey Methodology*, 22, 107-115.

STUKEL, D., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus linearization. *Survey Methodology* , 22, 117-125.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 21-30
Statistics Canada

Estimation in Sample Surveys Using Frames With a Many-to-Many Structure

TERRI L. BYCZKOWSKI, MARTIN S. LEVY and DENNIS J. SWEENEY[1]

ABSTRACT

In sample surveys, the units contained in the sampling frame ideally have a one-to-one correspondence with the elements in the target population under study. In many cases, however, the frame has a many-to-many structure. That is, a unit in the frame may be associated with multiple target population elements and a target population element may be associated with multiple frame units. Such was the case in a building characteristics survey in which the frame was a list of street addresses, but the target population was commercial buildings. The frame was messy because a street address corresponded either to a single building, multiple buildings, or part of a building. In this paper, we develop estimators and formulas for their variances in both simple and stratified random sampling designs when the frame has a many-to-many structure.

KEY WORDS: Imperfect frames; Correspondence errors; Building characteristics survey; Weighting; Simple random sampling; Stratified random sampling.

1. INTRODUCTION

This research was motivated by a study that was conducted for a utility company to estimate various population characteristics of the commercial buildings located in their service area. Budgetary constraints prohibited the development of a list of commercial buildings using canvassing techniques. However, a sampling frame consisting of street addresses (*i.e.*, addresses at which a utility meter was located) was available. A drawback of this frame was that it had a many-to-many relationship with the target population of commercial buildings. That is, some units in the frame were associated with multiple target population elements, and some target population elements were associated with multiple frame units. In fact, several of the relationships between street addresses and commercial buildings were relatively complex.

An advantage of this frame, however, was that total annual electrical usage was available for each street address. This resulted in a variable upon which the frame of street addresses could be effectively stratified. One of the important characteristics to be measured was the total commercial square footage. Studies conducted in the United States have shown that energy consumption is associated with both building size and building activity. For example, consumption is higher for buildings used for health care or food sales, and lower for buildings used for religious worship or public assembly. Also, energy consumption is correlated with building size even if the activity of the building is not known, as was the case here (U.S. Department of Energy 1992).

There is a vast amount of literature dealing with imperfect sampling frames. Comprehensive summaries of this literature can be found in Kish (1965), Wright and Tsao (1983), and Lessler and Kalsbeek (1992). Another body of literature addresses multiplicity sampling in which the frame is constructed with a many-to-many structure by design. Here, frame imperfections are introduced in order to gather information more efficiently on rare occurrences in a population (Birnbaum and Sirken 1965, Sirken 1972a,b, and Casady and Sirken 1980). Hansen, Hurwitz and Madow (1953a,b) present an estimator for use with sampling frames that have a many-to-one structure; population elements are represented multiple times in the frame. This estimator has also been adopted for use by National Agricultural Statistics Service (NASS) surveys (Musser 1993) with respect to the many-to-one frame. Bandyopadhyay and Adhikari (1993) developed estimators for a ratio, population mean, and population total when an unknown amount of duplication is present in the frame. But, these estimators are restricted to the simple random sampling case and the many-to-one frame.

Two methods for estimating population characteristics using a frame with a many-to-many structure appear in the literature. First, the Horvitz-Thompson estimator (1952) provides unbiased estimates of population means and totals when varying probabilities of selection are present. Musser (1993) shows how to compute the correct inclusion probabilities for the population elements selected in simple random sampling from a many-to-one frame. However, Musser's method can be extended to obtain inclusion probabilities for population elements in a simple random sample from the many-to-many frame as well. Second, Lavallée (1995) adapted the Weight Share Method, applied to longitudinal surveys, to the use of frames with a many-to-many structure.

The purpose of this paper is to develop an alternative methodology for estimating population totals, counts, and

[1] Terri L. Byczkowski, Institute for Policy Research, Martin S. Levy and Dennis J. Sweeney, Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, OH 45221, U.S.A.

means when using sampling frames with a many-to-many structure under simple and stratified random sampling designs. Also, expressions for the variance of those estimators are derived. The results which we develop are not only of intrinsic interest, but expressions for the variance of the estimators are essential for the exploration of the effects of correspondence imperfections inherent in many-to-many sampling frames on the precision of these estimates.

In section 2 we present these estimates in the simple random sampling without replacement (SRSWOR) case. We also describe the sampling methodology under which these estimators are applicable, state a result on bias, and develop expressions for their variance.

In section 3 some of the results are extended to the case of stratified random sampling. In section 4 we develop conclusions, discuss limitations and make suggestions for future research.

2. MANY-TO-MANY FRAMES FOR SIMPLE RANDOM SAMPLING

It is useful to think of the relationship between the frame and the target population as a graph. The sampling units in the frame and the elements of the target population are the two sets of nodes; arcs link the sampling units to elements of the target population. These arcs reveal the structure of the relationship between the frame and the target population. Figure 2.1 shows an example of a frame and target population with a many-to-many relationship. There are 7 sampling units in the frame, 6 elements in the target population and 10 links (arcs) between the sampling units and the elements of the population. Thus, a graph with 13 nodes and 10 arcs represents this many-to-many structure. In this paper we assume that each population element is linked to the set of frame units by at least one arc and that each frame unit is linked to the set of population elements by at least one arc as well.

Let us fix some notation. We find it convenient to identify both frame units and population elements with their respective indices. Let $F = \{1, 2, ..., N\}$ denote the set of indices for N sampling units, and let $T = \{1, 2, ..., M\}$ denote the set of indices for the M target population elements. An arc can be represented as an ordered pair; the first element of which comes from F, and the second from T. A population element k in T is said to be represented by sampling unit j in F, if it is linked to it by an arc denoted (jk). This means that when j is in the sample there is a nonzero probability of collecting data from population element k. We will denote by y_k the measurement of interest on target population element k in T.

We now describe the sampling methodology under which the estimators developed herein are appropriate. Assume a SRSWOR of size n frame units is selected from F. The number of *population elements* included in the sample and measured, however, depends upon the nature of the association between the frame units and the population elements.

Under SRSWOR, one of four scenarios can occur when a frame unit is selected. In the first scenario, a frame unit corresponds to one and only one population element (a one-to-one structure). Here the surveyor would simply collect the information concerning the single population element corresponding to the selected frame unit (see frame unit 1 of Figure 2.1).

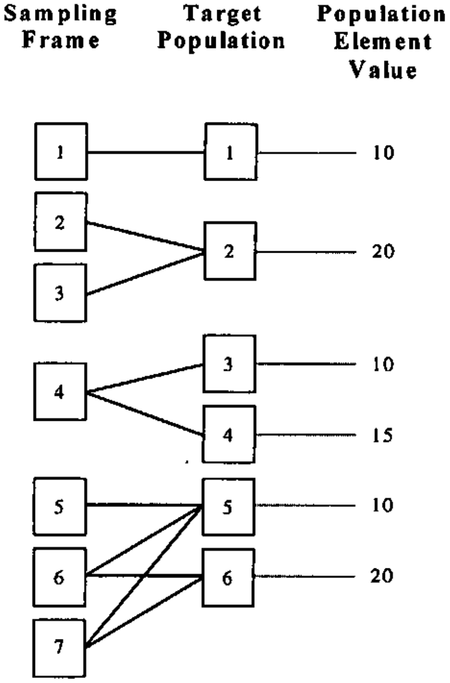


Figure 2.1. An example of the correspondence between the sampling frame and the target population

In the second scenario, several frame units correspond to one population element (a many-to-one structure). For example, in Figure 2.1, frame units 2 and 3 correspond to the single population element 2. In this case, if frame units 2 and/or 3 are included in the sample, information on population element 2 is collected. Thus, it is possible that population element 2 could appear in the sample, and as a record in the data set used to develop the estimates, up to two times.

In the third scenario, one frame unit corresponds to more than one population element (a one-to-many structure). For example, in Figure 2.1 frame unit 4 corresponds to population elements 3 and 4. Here, only one population element (3 or 4) is selected using a *randomization* independent of the choice of frame units. Economics dictated this policy because data collection entailed lengthy personal interviews conducted by individuals with technical backgrounds. In this paper we assume that these randomizations are conducted using equal probabilities. But, any probabilities could be used (*e.g.*, probability proportional to size) provided they are non-zero.

In the fourth scenario, a many-to-many structure exists. This is illustrated by frame units 5, 6 and 7 and population elements 5 and 6 in Figure 2.1. Since these complex cases are combinations of scenarios 2 and 3 above, the same sampling rules apply. For example, if frame unit 5 is selected, population element 5 is measured. If frame unit 6 is selected, only one of population elements 5 and 6 is randomly selected and measured.

2.1 Population Totals

2.1.1 Estimator for a Population Total

A many-to-many frame results in varying probabilities of selection. The estimators developed here involve a method of weighting, which is an extension of the estimator presented by Hansen *et al.* (1953a pp. 62-64). Their estimators and formulas for the variance of those estimators are restricted to the many-to-one frame structure. We extend those estimators to the many-to-many frame structure.

For a SRSWOR of size n, let $J_1, ..., J_n$ denote random variables such that $J_i = j$ if the i-th draw results in the selection of unit j from F. Hence $\Pr(J_i = j) = 1/N$ for j in F and $i = 1, ..., n$. Let $K_1, ..., K_n$ denote random variables such that $K_i = k$ if the i-th draw from F is followed by the selection of k from T. We can now think of drawing a random sample of arcs $\{(J_1K_1), ..., (J_nK_n)\}$ which has a joint probability distribution determined by both the SRSWOR sampling design and the subsequent randomization (if required) to choose an element in T. In particular, (J_iK_i) has marginal probability given by $\Pr\{(J_iK_i) = (jk)\} = (1/N)s_{jk}$, in which s_{jk} is the conditional probability given by, $s_{jk} = \Pr(K_i = k \mid J_i = j)$. That is, s_{jk} is the conditional probability of selecting population element k in T given that frame unit j in F is selected. These conditional probabilities will be referred to as arc probabilities and are illustrated for Figure 2.1 in Table 2.1.

Table 2.1
Arc Probabilities for Figure 2.1

Arc jk	1,1	2,2	3,2	4,3	4,4	5,5	6,5	6,6	7,5	7,6
s_{jk}	1	1	1	1/2	1/2	1	1/2	1/2	1/2	1/2

For k in T, let U_k denote the set of units in F that have arcs with a destination at k in T. Let $s_k = \sum_{j\in U_k} s_{jk}$. Using the language in Hansen *et al.* (1953a pp. 62-64) which motivated our development, we call s_k the *weight* for population element k in T. These weights for Figure 2.1 appear in Table 2.2.

Table 2.2
Calculation of the Population Element Weights (s_k) for Figure 2.1

k	1	2	3	4	5	6
(s_k)	1	2	1/2	1/2	2	1

Arc probabilities and weights are used to compute the marginal probabilities of the K_i, namely, $\Pr(K_i = k) = \sum_{j\in U_k}(1/N)s_{jk} = (1/N)s_k$, where k is in T, and $i = 1, ..., n$. Clearly, computing the arc probabilities is the key step in developing the correct weights for the data collected. It depends on properly ascertaining the graph structure for each sampling unit selected: a maximally connected (MC) subgraph. A connected subgraph is a subset of the nodes which are connected by a sequence of arcs. Maximal means that no node outside the subset is connected to a node belonging to the subset. There are 4 MC subgraphs in Figure 2.1. Each represents a different frame – population structure, namely, one-to-one, many-to-one, one-to-many, and many-to-many structure.

To develop the estimators it is not necessary to know the structure for the entire graph. It is only necessary to know the structure of the MC subgraphs to which *sampled* frame units belong.

We make the following observations about s_k and s_{jk}: (i) $s_k = W$ indicates that population element k has W times the probability of being selected on the i-th draw as that of a population element with a weight of one; (ii) $0 < s_k \le N$, $k = 1, ..., M$; (iii) $0 < s_{jk} \le 1$, $j \in U_k$ and $k = 1, ..., M$; (iv) with respect to the one-to-many frame structure, $s_{jk} = s_k$; (v) with respect to the many-to-one frame structure, $s_{jk} = 1$ for all k; and (vi) $\sum_{k=1}^{M}\sum_{j=1}^{N} s_{jk} = N$.

Now, let $x_1, ..., x_M$ denote the weighted values associated with the indices in T. That is, let $x_k = y_k/s_k$. Define random variables $x_{K_1}, ..., x_{K_n}$, associated with draws 1 through n from F, respectively, so that x_{K_i} takes the value x_k if $K_i = k$. Notice that we can write,

$$E(x_{K_i}) = \sum_{k=1}^{M} x_k \Pr(K_i = k) = \frac{1}{N}\sum_{k=1}^{M}\frac{y_k}{s_k}s_k = \frac{Y}{N}, \qquad (2.1)$$

where $Y = \sum_{k=1}^{M} y_k$ is the true population total. We take as our estimator of the population total based upon a SRSWOR from a sampling frame with many-to-many structure,

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} x_{K_i}. \qquad (2.2)$$

Using (2.1) it follows that,

$$E(\hat{Y}) = E\left(\frac{N}{n}\sum_{i=1}^{n} x_{K_i}\right) = \frac{N}{n}\sum_{i=1}^{n} E(x_{K_i}) = \frac{N}{n}n\frac{Y}{N} = Y.$$

We thus obtain,

Theorem 2-1: The estimator (2.2) for a population total used in SRSWOR is unbiased.

Using Figure 2.1, we now give a simple example of the use of this estimator. Suppose a simple random sample of four frame units was selected from the frame depicted in Figure 2.1 (2, 3, 4, and 7) which ultimately resulted in the selection of population elements 2, 4, and 5. The estimator of the population total,

$\hat{Y} = \frac{N}{n}\sum_{i=1}^{4} x_{K_i}$, has value $\frac{7}{4}\left[\frac{20}{2}+\frac{20}{2}+\frac{15}{(1/2)}+\frac{10}{2}\right]=\frac{385}{4}$.

The above estimator can also be used for a population count. We could estimate the size of the target population by letting $y_k = 1$ for all k. In addition, we could estimate the number of population elements that possess some characteristic by letting $y_k = 1$ for those population elements with the characteristic of interest and $y_k = 0$ for those without the characteristic.

2.1.2 Variance of the Estimator for a Population Total

First, some additional terminology and notation used in this section must be defined. Let P represent the set of all unordered pairs of arcs. We shall define an unordered pair of arcs as *inadmissible* if they cannot both be included in a sample. Formally let $Q = \{ j \text{ in } F\text{: more than one arc emerges from } j \}$. Then $R' = \{[jk, jk']: j \in Q \text{ and } k \neq k'\}$ is the set of unordered *inadmissible* pairs of arcs. Also, the set of unordered *admissible* pairs of arcs is the complementary set $R^* = P \setminus R'$.

To illustrate, consider Figure 2.1. The sampling methodology we employ requires that if frame unit 4 is selected, only one of population elements 3 and 4 can be included in the sample. Thus, {[4,3][4,4]} is an unordered inadmissible pair of arcs. The other unordered inadmissible pairs of arcs in Figure 2.1 are {[6,5][6,6]} and {[7,5][7,6]}. Thus, R' = {[4,3][4,4], [6,5][6,6], [7,5][7,6]}.

Theorem 2-2: The variance of the estimator (2.2) is,

$$V(\hat{Y}) = \frac{N}{n}\left[\sum_{k=1}^{M}\frac{y_k^2}{s_k} + 2\frac{(n-1)}{(N-1)}\sum \sum_{[jk,j'k']\in R^*}\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right)\right] - Y^2, \quad (2.3)$$

where the double sum is over all unordered *admissible* pairs of arcs $[jk, j'k']$.

Proof:

$$V(\hat{Y}) = E\left[\left(\frac{N}{n}\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2$$

$$= \frac{N^2}{n^2}E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2. \quad (2.4)$$

Now,

$$E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] = \sum_{i=1}^{n} E\left(x_{K_i}^2\right) + 2E\left(\sum\sum_{i<i'}\left(x_{K_i}x_{K_{i'}}\right)\right). \quad (2.5)$$

One can write

$$E\left(x_{K_i}^2\right) = \sum_{k=1}^{M}\left[x_k^2 \Pr(K_i = k)\right] = \sum_{k=1}^{M}\frac{y_k^2}{s_k^2}\frac{s_k}{N} = \frac{1}{N}\sum_{k=1}^{M}\frac{y_k^2}{s_k}. \quad (2.6)$$

As mentioned in Section 2.1, we can think of selecting a sample of arcs which ultimately leads to the selection of population elements. Each arc (jk) is associated with a value $x_k = y_k/s_k$ of the population element k at its destination. Thus, we can rewrite the double summation in (2.5) as a summation over admissible unordered pairs of arcs, R^*.

$$2E\left(\sum_{i'}\sum_{i<i'}\left(x_{K_i}x_{K_{i'}}\right)\right) =$$

$$2\binom{n}{2}\sum\sum_{[jk,j'k']\in R^*}\left[(x_k x_{k'})\Pr(K_i = k, K_{i'} = k')\right]. \quad (2.7)$$

Now, by virtue of the independence of the randomization and the choice of frame units:

$$\Pr(\text{select}[jk, j'k'] \text{ in } R^*) = \Pr(\text{select}\{j, j'\} \text{ in } F)$$

$$\Pr(\text{select}[jk, j'k'] \text{ in } R^* \mid \text{select}\{j, j'\} \text{ in } F) = \frac{1}{\binom{N}{2}} s_{jk}s_{j'k'}.$$

Substituting into (2.7) results in,

$$n(n-1)\sum\sum_{[jk,j'k']\in R^*}\left[(x_k x_{k'})\frac{1}{\binom{N}{2}}s_{jk}s_{j'k'}\right] =$$

$$\frac{2n(n-1)}{N(N-1)}\sum\sum_{[jk,j'k']\in R^*}\left[\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right)\right]. \quad (2.8)$$

Now substituting (2.6) and (2.8) into (2.5) yields,

$$E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] = \frac{n}{N}\left[\sum_{k=1}^{M}\frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*}\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right)\right]. \quad (2.9)$$

Finally substituting (2.9) into (2.4) gives the result (2.3).

Equation (2.3) is a generalization of the formula developed by Bandyopadhyay and Adhikari (1993) for the variance of the estimate of a population total in the case of the many-to-many frame structure. It can be shown that (2.3) reduces to their formula when the sampling frame is restricted to a many-to-one structure.

Corollary 2-1: An alternative form of the variance formula in **Theorem 2-2** is:

$$\mathrm{V}(\hat{Y}) = \frac{N}{n}\left[\sum_{k=1}^{M}\frac{y_k^2}{s_k} + \frac{(n-1)}{(N-1)}\left(\left(\sum_{jk}\frac{y_k}{s_k}s_{jk}\right)^2 - \sum_{jk}\left(\frac{y_k}{s_k}s_{jk}\right)^2 - 2\sum\sum_{[jk,jk']\in R'}\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}\right)\right] - Y^2.$$

Proof:
Write,

$$\left(\sum_{jk}\frac{y_k s_{jk}}{s_k}\right)^2 = \sum_{jk}\left(\frac{y_k s_{jk}}{s_k}\right)^2 + 2\sum_{[jk,j'k']\in R^*}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}} + 2\sum_{[jk,jk']\in R'}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}.$$

It follows that:

$$\sum_{[jk,j'k']\in R^*}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}} = \frac{1}{2}\left(\sum_{jk}\frac{y_k s_{jk}}{s_k}\right)^2 - \frac{1}{2}\sum_{jk}\left(\frac{y_k s_{jk}}{s_k}\right)^2 - \sum_{[jk,jk']\in R'}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}.$$

Substituting the above expression into (2.3) provides the result.

This formula is computationally simpler. Note that (2.3) requires that the term

$$\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right)$$

be summed over all unordered admissible pairs of arcs (R^*), whereas this alternative formula only requires a summation over pairs of arcs that are inadmissible (R'). In most practical scenarios the number of admissible pairs of arcs will be far greater than the number of inadmissible pairs of arcs.

2.2 Population Means

2.2.1 Estimator for a Population Mean

The estimator for a population mean presented here extends the estimator presented by Hansen *et al.* (1953a) to the many-to-many frame structure.

Associated with the n draws from F, define random variables s_{K_i} and $z_{K_i} = 1/s_{K_i}$, so that s_{K_i} takes value s_k if $K_i = k$ for $i = 1, ..., n$ and $k = 1, ..., M$. The estimator for a population mean,

$$\bar{Y} = \frac{1}{M}\sum_{k=1}^{M} y_k,$$

when using SRSWOR and a many-to-many frame is:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} x_{K_i}}{\sum_{i=1}^{n} z_{K_i}}. \qquad (2.10)$$

2.2.2 Mean Square Error (MSE) of the Estimator for a Population Mean

The estimator for a population mean is biased because it is a ratio estimator. But, it is well known that this bias becomes negligible for large samples and the bias is of order $1/n$ (Cochran 1977, p. 160).

Our approximation of the MSE requires a summation over R^{**}, the set of all *ordered admissible* pairs of arcs. Thus, if $[jk,j'k']\in R^*$, then both $[jk,j'k']\in R^{**}$ and $[j'k',jk]\in R^{**}$.

To approximate the mean square error of the estimator (2.10), we use

$$\mathrm{MSE}(\hat{\bar{Y}}) \approx \frac{M^2}{nN\left(\sum_{k=1}^{M}\frac{1}{s_k}\right)^2}\left[\left(\sum_{k=1}^{M}\frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*}\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right) - 2\bar{Y}\left(\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^{**}}\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right) + \bar{Y}^2\left(\sum_{k=1}^{M}\frac{1}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*}\frac{s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right)\right]. \qquad (2.11)$$

To justify this approximation let

$$\bar{x} = \frac{\sum_{i=1}^{n} x_{K_i}}{n}, \quad \bar{z} = \frac{\sum_{i=1}^{n} z_{K_i}}{n} \text{ and } \bar{Z} = \frac{\sum_{k=1}^{M}\frac{1}{s_k}}{M}.$$

Because $\hat{\bar{Y}}$ is a ratio of two estimates, the well known approximation for the mean square error (Cochran 1977, pp. 32-33) can be used:

$$\mathrm{MSE}(\hat{\bar{Y}}) = \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{z}}\right)^2 \approx \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{Z}}\right)^2 =$$

$$\frac{1}{\bar{Z}^2}\left[\mathrm{E}(\bar{x}^2) - 2\bar{Y}\mathrm{E}(\bar{z}\bar{x}) + \bar{Y}^2\mathrm{E}(\bar{z}^2)\right] =$$

$$\frac{M^2}{n^2\left(\sum_{j=1}^{M} z_j\right)^2}\left[\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\right)^2 -\right.$$

$$\left.2\bar{Y}\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\sum_{i=1}^{n} z_{K_i}\right) + \bar{Y}^2\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2\right]. \tag{2.12}$$

The first expectation in (2.12) is simply (2.9). Next, using (2.1) on the middle term in (2.12) results in

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\sum_{i=1}^{n} z_{K_i}\right) = \mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\frac{1}{s_{K_i}}\right) +$$

$$\mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i'}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) = \frac{n}{N}\sum_{k=1}^{M}\frac{y_k}{s_k} + \mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right).$$

Using (2.7) and (2.9) yields,

$$\mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) = n(n-1)\mathrm{E}\left(x_{K_i}\frac{1}{s_{K_{i'}}}\right) =$$

$$n(n-1)\sum_{[jk,j'k']\in R^{**}}\sum\left[\left(\frac{y_k}{s_k}\frac{1}{s_{k'}}\right)\Pr\left(x_{K_i} = \frac{y_k}{s_k}, \frac{1}{s_{K_{i'}}} = \frac{1}{s_{k'}}\right)\right] =$$

$$n(n-1)\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k}{s_k}\frac{1}{s_{k'}}\left(\frac{1}{N(N-1)}s_{jk}s_{j'k'}\right) =$$

$$\frac{n(n-1)}{N(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}.$$

Note that the double sum is over all admissible *ordered* pairs of arcs. Therefore,

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\frac{1}{s_{K_i}}\right) + \mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i'}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) =$$

$$\frac{n}{N}\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{n(n-1)}{N(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}} =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{(n-1)}{(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Finally, similar to (2.1),

$$\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}\frac{1}{s_{K_i}}\right)^2 =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{1}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^{*}}\frac{S_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Substituting these expectations into equation (2.12) yields (2.11).

3. ESTIMATORS FOR MANY-TO-MANY FRAMES UNDER STRATIFIED RANDOM SAMPLING

3.1 Introduction

In this section we develop the estimators for a population count, mean, and total in the many-to-many frame case, when stratified random sampling is used. First, however, it is necessary to describe the sampling methodology under which these estimates are appropriate. Figure 3.1 provides an example that will be used throughout this section.

3.2 The Sampling Methodology

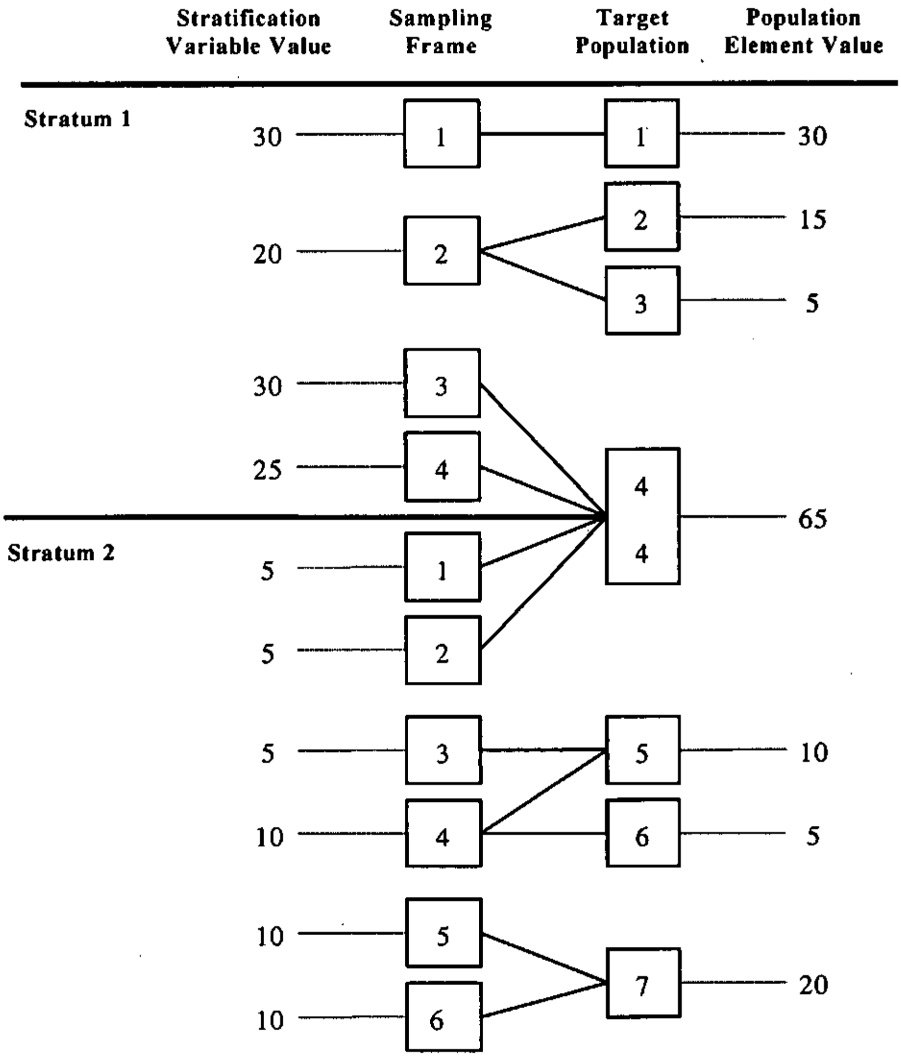
The same scenarios that were described in SRSWOR occur with respect to stratified random sampling. However, there are some additional problems that can arise in this case.

Consider the building characteristics study that motivated this research. Assume that the population element value in Figure 3.1 is the building size, and the stratification variable is electrical usage associated with the street address. Because the frame of street addresses had a many-to-many correspondence with the target population of commercial buildings, the following problems arose in addition to those mentioned in Section 2.1:

1. Mis-stratification: For example, frame unit (street address) 2 in stratum 1 appeared to be a large building because of the large electrical usage associated with it, and as a result, it was placed in the first stratum. The data collection revealed that the street address actually corresponded to two small buildings (population elements 2 and 3). In another example, frame units 5 and 6 in stratum 2 appeared to be two small buildings in the frame, and were placed in the second stratum. But, the corresponding population element 7 is one large building with two street addresses.
2. Crossover: For example, frame units 3 and 4 in stratum 1, and frame units 1 and 2 in stratum 2 each have a different street address and, as a result, appear in the frame to be two small and two large buildings. But, data

collection revealed that all four street addresses corresponded to only one building (*e.g.*, a strip mall). In this case, not only is mis-stratification a problem, but not all the frame units associated with a single building are included in the same strata. That is, one population element (*i.e.*, building) "crosses over" multiple strata.

In the next section we develop estimators for population totals and counts and show that these estimators are unbiased despite mis-stratification and crossover. As is usually the case, however, mis-stratification increases the variance of the estimates. Also, insofar as crossover induces mis-stratification, it too increases the variances of the estimates.



Note: Frame units were placed in stratum 1 if the value of the stratification variable was 20 or more. Otherwise, the frame units were placed in stratum 2.

Figure 3.1. An example of the correspondence between the frame and the target population in stratified random sampling

Table 3.1
Arc Probabilities for Figure 3.1

Arc hjk	1,1,1	1,2,2	1,2,3	1,3,4	1,4,4	2,1,4	2,2,4	2,3,5	2,4,5	2,4,6	2,5,7	2,6,7
s_{hjk}	1	1/2	1/2	1	1	1	1	1	1/2	1/2	1	1

3.3 Population Totals and Counts

3.3.1 Estimator for a Population Total

The estimator developed here involves a method of weighting which extends the estimator presented in Hansen *et al.* (1953a, pp. 62-64) to stratified random sampling when using a many-to-many frame.

Assume that F has been partitioned into L mutually exclusive and exhaustive strata $F_1, ..., F_L$ of size $N_1, ..., N_L$ respectively. Units in F_h will be denoted hj where $j = 1, ..., N_h$ and $h = 1, ..., L$. Also, assume that a stratified random sample (without replacement) of size $n = n_1 + ... + n_L$ has been drawn, where n_h is the sample size from F_h. Let $hJ_1, ..., hJ_{n_h}$ denote random variables such that $hJ_i = hj$ if the i-th draw from F_h results in the selection of hj. Let $hK_1, ..., hK_{n_h}$ denote random variables such that $hK_i = k$ if the i-th draw from F_h is followed by the selection of k from T. If hjk denotes the arc that originates at frame unit hj in F_h and terminates at k in T, the marginal probability of the random arc (hJ_i, hK_i) is given by,

$$\Pr\{(hJ_i, hK_i) = (hjk)\} = \frac{1}{N_h} s_{hjk},$$

in which $s_{hjk} = \Pr(hK_i = k \mid hJ_i = hj)$ is an arc probability. Note that s_{hjk} is the conditional probability of selecting population element k in T given that frame unit hj has been chosen. Assuming equal randomization probabilities, Table 3.1 shows the arc probabilities for Figure 3.1.

Let W_k denote the set of frame units hj in F that have arcs with a destination at k in T. For example, $W_4 = \{(1, 3), (1, 4), (2, 1), (2, 2)\}$. Also, define the population element weight $s_k = \sum_{hj \in W_k} s_{hjk}$.

Table 3.2 contains the weights (s_k) for all the population elements in Figure 3.1. The same observations concerning arc weights (s_{hjk}) and population element weights (s_k) made in section 2.3.1 apply here.

Table 3.2
Population Element Weights (s_k) for Figure 3.1

k	1	2	3	4	5	6	7
s_k	1	1/2	1/2	1+1+1+1=4	1+1/2=3/2	1/2	1+1=2

For each $h = 1, ..., L$ and $i = 1, ..., n_h$, let x_{hK_i} be random variables such that $x_{hK_i} = y_k/s_k$ if k in T is selected as a result of the selection of some hj in F_h.

The estimator of a population total for stratified random sampling, when using a sampling frame with a many-to-many structure is:

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h, \text{ where } \hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hK_i}. \qquad (3.1)$$

3.3.2 Variance of the Estimator for a Population Total

Prior to developing the variance of estimator (3.1), some additional terminology must be defined. Let q_{hk} denote the "stratum element weight". This additional weight is necessary because of the potential of crossover. Let U_{hk} denote the set of frame units in F_h that have arcs with a destination at population element k. For example, $U_{24} = \{(2, 1), (2, 2)\}$. Then define $q_{hk} = \sum_{hj \in U_{hk}} s_{hjk}$. To illustrate, recall in Figure 3.1 population element 4 is represented by two frame units in stratum 2, so $q_{24} = \sum_{2j \in U_{24}} s_{2j4} = 2$.

The weight q_{hk} plays the role of s_k when selection is restricted to F_h. In fact, $q_{hk} = s_k$ when there is no crossover. The probability of selecting any frame unit from F_h on step i out of n_h is $1/N_h$. But, the probability of selecting a population element k represented by a frame unit in F_h is $\Pr(hK_i = k) = q_{hk}/N_h$, for all $i = 1, ..., n_h$.

In order to develop the proof in this section, we introduce the term "apportioned stratum total" denoted by Y_h^*. In effect, the values of the population elements that are represented by frame units in multiple strata are apportioned among those strata. Let V_h denote the set of population elements associated with frame units in F_h. In our example $V_1 = \{1, 2, 3, 4\}$ and $V_2 = \{4, 5, 6, 7\}$. Let

$$Y_h^* = \sum_{k \in V_h} y_k q_{hk} / s_k$$

where y_k is the value of population element k, $k = 1, 2, ..., M$. When crossover is present, use of the weights q_{hk} and s_k apportion the measure y_k among the strata in which population element k is represented. We can think of the use of these weights as distributing the population element value among the strata depending upon the number of times the population element is represented in a stratum relative to the total number of times it is represented in the frame. For example in Figure 3.1 Y_1^* and Y_2^* are calculated as follows:

$$Y_1^* = \frac{30(1)}{1} + \frac{15(1/2)}{1/2} + \frac{5(1/2)}{1/2} + \frac{65(2)}{4} = 82.5$$

$$Y_2^* = \frac{65(2)}{4} + \frac{10(3/2)}{3/2} + \frac{5(1/2)}{1/2} + \frac{20(2)}{2} = 67.5.$$

Note that $\sum_{h=1}^{L} Y_h^* = Y$ whether or not crossover exists.

Theorem 3-1: The estimator for a population total (3.1) is unbiased.

Proof:

From (3.1),

$$\mathrm{E}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \mathrm{E}\left(x_{hK_i}\right). \qquad (3.2)$$

For each $i = 1, ..., n_h$,

$$\mathrm{E}\left(x_{hK_i}\right) = \sum_{k \in V_h} \frac{y_k}{s_k} \Pr(hK_i = k) =$$

$$\sum_{k \in V_h} \frac{y_k}{s_k} \frac{q_{hk}}{N_h} = \frac{1}{N_h} \sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} = \frac{1}{N_h} Y_h^*. \qquad (3.3)$$

Substituting (3.3) into equation (3.2) yields $E(\hat{Y}_{st}) = Y$.

In the main result below we need the following notation. Let R_h^* and R_h' be the set of admissible and inadmissible unordered pairs of arcs originating in F_h, respectively. Definitions of the above are identical to the corresponding concepts for the SRSWOR case, but restricted now to strata.

Theorem 3-2: The variance of (3.1) is:

$$\mathrm{V}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} S_h^2, \tag{3.4}$$

where,

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k\in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{2(n_h - 1)}{(N_h - 1)} \times \right.$$
$$\left. \sum_{[hjk,hj'k']\in R_h^*}\left(\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hj'k'}}{s_{k'}}\right)\right] - \left(\sum_{k\in V_h}\frac{y_k q_{hk}}{s_k}\right)^2. \tag{3.5}$$

Proof: First write,

$$\mathrm{V}\left(\hat{Y}_{st}\right) = \mathrm{E}\left(\hat{Y}_{st}\right)^2 - Y^2 = \mathrm{E}\left(\sum_{h=1}^{L}\hat{Y}_h^2\right) -$$
$$\sum_{h=1}^{L}\left(Y_h^*\right)^2 + 2\left(\mathrm{E}\left(\sum_{h<h'}\hat{Y}_h\hat{Y}_{h'}\right) - \sum_{h<h'} Y_h^* Y_{h'}^*\right). \tag{3.6}$$

The last two terms cancel because $\hat{Y}_h$ and $\hat{Y}_{h'}$ are independent. This follows since apportionment creates a new stratified population containing no crossover and samples chosen within different strata are independent. Thus, with

$$S_h^2 = \mathrm{E}\left(\hat{Y}_h\right)^2 - \left(Y_h^*\right)^2, \quad \mathrm{V}\left(\hat{Y}_{st}\right) = \mathrm{E}\left(\sum_{h=1}^{L}\hat{Y}_h^2\right) -$$
$$\sum_{h=1}^{L}\left(Y_h^*\right)^2 = \sum_{h=1}^{L}\left(\mathrm{E}\left(\hat{Y}_h^2\right) - \left(Y_h^*\right)^2\right) = \sum_{h=1}^{L} S_h^2.$$

Now,

$$\mathrm{E}\left(\hat{Y}_h\right)^2 = \frac{N_h^2}{n_h^2}\mathrm{E}\left(\sum_{i=1}^{n_h} x_{hK_i}\right) =$$
$$\frac{N_h^2}{n_h^2}\left(\sum_{i=1}^{n_h}\mathrm{E}\left(x_{hK_i}\right)^2 + 2\mathrm{E}\left(\sum_{i<i'} x_{hK_i}x_{hK_{i'}}\right)\right). \tag{3.7}$$

For each $i = 1, \ldots, n_h$,

$$\mathrm{E}\left(x_{hK_i}\right)^2 =$$
$$\sum_{k\in V_h}\left[\left(\frac{y_k}{s_k}\right)^2 \Pr\left(hK_i = k\right)\right] = \sum_{k\in V_h}\left[\left(\frac{y_k}{s_k}\right)^2\frac{q_{hk}}{N_h}\right]. \tag{3.8}$$

Then, using equation (2.7) and (2.8),

$$2\mathrm{E}\left(\sum_{i>i'} x_{hK_i}x_{hK_{i'}}\right) = 2\binom{n_h}{2}\mathrm{E}\left(x_{hK_i}x_{hK_{i'}}\right) =$$
$$n_h(n_h - 1)\sum_{[hjk,hj'k']\in R_h^*}\frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\Pr(hK_i = k, hK_{i'} = k') =$$
$$n_h(n_h - 1)\sum\sum_{[hjk,hj'k']\in R_h^*}\left[\left(\frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\right)\binom{N_h}{2}^{-1} s_{jk}s_{j'k'}\right] =$$
$$\frac{2n_h(n_h - 1)}{N_h(N_h - 1)}\sum\sum_{[hjk,hj'k']\in R_h^*}\left[\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hj'k'}}{s_{k'}}\right]. \tag{3.9}$$

Equation (3.5) now follows from (3.8), (3.9), and the definition of Y_h^*.

Using the method of Corollary 2-1, (3.5) can be simplified for computing purposes as follows:

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k\in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{(n_h - 1)}{(N_h - 1)}\left[\left(\sum_{hjk\in A_h}\frac{y_k s_{hjk}}{s_k}\right)^2 - \right.\right.$$
$$\left.\left.\sum_{hjk\in A_h}\left(\frac{y_k s_{hjk}}{s_k}\right)^2 - 2\sum_{[hjk,hjk']\in R_h'}\left(\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hjk'}}{s_{k'}}\right)\right]\right] -$$
$$\left(\sum_{k\in V_h}\frac{y_k q_{hk}}{s_k}\right)^2,$$

where A_h denotes the set of arcs that originate at frame units in F_h.

3.4 Population Means

3.4.1 Estimator for a Population Mean

The estimator developed here for a population mean for stratified random sampling extends the estimator presented by Hansen *et al.* 1953a (pp. 62-64) to the case of a stratified random sample from a many-to-many frame.

The estimator for a population mean when using stratified random sampling and a many-to-many frame is:

$$\hat{\hat{Y}}_{st} = \sum_{h=1}^{L}\frac{N_h}{N}\hat{\hat{Y}}_h, \text{ where } \hat{\hat{Y}}_h = \frac{\sum_{i=1}^{n_h} x_{hK_i}}{\sum_{i=1}^{n_h}\frac{1}{s_{hK_i}}}. \tag{3.10}$$

As in the SRSWOR case, the estimator for a population mean is biased because it is a ratio estimator.

4. CONCLUSIONS

In this paper we have developed estimators for population totals, counts and means that are appropriate when the sampling frame has a many-to-many structure. We have focused on simple random sampling and stratified random sampling designs.

We used the method of weighting described in this paper in a study of commercial buildings for which a stratified random sample was employed. In this study, for which the sampling frame consisted of street addresses, interviewers recorded any additional street addresses that pertained to the selected building. It was then determined whether or not these additional street addresses were listed in the sampling frame, and whether or not they were connected to other population elements (commercial buildings). In more complex scenarios, the interviewers sometimes resorted to schematic sketches of the buildings and labelling all the pertinent addresses. This allowed us to determine the structure of all MC subgraphs in our sample and to develop the appropriate weights s_k.

In addition, we developed formulas for the variance of some of the estimators presented in this paper. It should be noted that these variance formulas are population parameters and do not translate readily into corresponding sample estimates. In fact, the authors are unaware of any optimal method for estimating the variances discussed in this paper. However, there are many computer intensive methods (balanced repeated replication, bootstrapping, *etc.*) for estimating variances in complex sample surveys (Wolter 1985). It should be emphasized that when using our estimators, each of these variance estimation schemes aims at a common target: the variance formulas we have developed.

Nevertheless, the usefulness of these variance formulas is in their application to the task of exploring the effects of frame imperfections, along with population characteristics, on the precision of estimation. Such an exploration, another future area of research, should result in recommendations and guidelines for the survey researcher on how to manage a frame with a many-to-many structure. That is, based upon frame and population characteristics, the survey researcher would be able to make strategic decisions concerning the options available: canvassing a population to remove correspondence imperfections, or using the estimators described herein.

Another area of future research is a comparison of the precision of our estimators to that of other estimators, such as the Horvitz-Thompson estimator. As noted in the introduction the Horvitz-Thompson estimator can be applied to sampling involving a many-to-many frame structure. An advantage of the Horvitz-Thompson estimator is that with properly identified first and second order inclusion probabilities, one can obtain both an estimate of a population characteristic and an unbiased estimate of its variance. In addition, the first order inclusion probabilities can be derived in a manner similar to Musser (1993) based only upon information from the MC subgraphs. However, these probabilities are very difficult to compute in a complex many-to-many frame structure such as ours. It is, however, relatively easy to calculate the necessary weights for our estimators.

ACKNOWLEDGMENTS

The authors thank the individuals at Cinergy Corporation for the opportunity to conduct the building characteristics survey which motivated this research. We also thank the three referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

BANDYOPADHYAY, S., and ADHIKARI, A.K. (1993). Sampling from imperfect frames with unknown amount of duplication. *Survey Methodology*, 19, 193-197.

BIRNBAUM, Z.W., and SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, PHS Publication 1000, Ser. 2, *Data Evaluation and Methods Research*, no. 11. Hyattsville, MD: National Center for Health Statistics, Public Health Service, U.S. Department of Health and Human Services.

CASADY, R.J., and SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-605.

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

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

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953b). *Sample Survey Methods and Theory 2, Theory*. New York: Wiley & Sons.

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.

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

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

LESSLER, J.T., and KALSBEEK, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley & Sons.

MUSSER, O. (1993). Unbiased estimation in the presence of frame duplication. *Proceedings of the International Conference on Establishment Surveys*, 889-892.

SIRKEN, M.G. (1972a). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

SIRKEN, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 65, 224-227.

U.S. DEPARTMENT OF ENERGY, Energy Information Administration (1992). *Commercial Buildings Energy Consumption Survey*.

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

WRIGHT, T., and TSAO, H.J. (1983). A frame on frames: An annotated bibliography, (Ed., Tommy Wright). *Statistical Methods and the Improvement of Data Quality*, Orlando, Florida: Academic Press, 25-72.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 31-40
Statistics Canada

Optimal Recursive Estimation for Repeated Surveys

IBRAHIM S. YANSANEH and WAYNE A. FULLER[1]

ABSTRACT

Least squares estimation for repeated surveys is addressed. Several estimators of current level, change in level and average level for multiple time periods are developed. The Recursive Regression Estimator, a recursive computational form of the best linear unbiased estimator based on all periods of the survey, is presented. It is shown that the recursive regression procedure converges; and that the dimension of the estimation problem is bounded as the number of periods increases indefinitely. The recursive procedure offers a solution to the problem of computational complexity associated with minimum variance unbiased estimation in repeated surveys. Data from the U.S. Current Population Survey are used to compare alternative estimators under two types of rotation designs: the intermittent rotation design used in the U.S. Current Population Survey, and two continuous rotation designs.

KEY WORDS: Recursive regression estimation; Composite estimation; Rotation designs; Rotation groups.

1. INTRODUCTION

We consider least squares estimation for surveys conducted on repeated occasions with partial overlap of sampling units. See Duncan and Kalton (1987) for a general discussion of different types of surveys and the objectives of such surveys. In this paper, we shall be concerned with rotating panel surveys, in which repeated determinations are made on some sampling units but not every unit appears in the sample at every time point.

Theoretical foundations for the design and estimation for repeated surveys based on generalized least squares procedures were laid down by Patterson (1950), following initial work by Cochran (1942) and Jessen (1942). Least squares procedures have been considered further by several other authors. See, for instance, Fuller (1990), and the references cited therein. Least squares estimation for a fairly general class of repeated surveys was considered by Yansaneh (1992). Composite estimation is a procedure of estimation for repeated surveys which makes use of the observations from the current and preceding periods, and the estimator of level from the preceding period. Breau and Ernst (1983) compared various alternative estimators to a composite estimator for the U.S. Current Population Survey (CPS). Kumar and Lee (1983) did similar work using data from the Canadian Labor Force Survey (LFS). Wolter (1979) provided a general composite estimation strategy for two-level rotation schemes such as the one used in the U.S. Census Bureau's Retail Trade Survey. Singh (1996) has proposed an alternative class of composite estimators. These authors assumed the unknown quantities on each occasion to be fixed parameters. Other authors, such as Scott, Smith, and Jones (1977), Jones (1980), Binder and Dick (1989), Bell and Hillmer (1990), and Pfeffermann (1991) considered estimation for repeated surveys under the assumption that the underlying true values constitute a realization of a time series.

In this paper, we discuss estimation procedures for repeated surveys, under the assumption that the unknown true values are fixed parameters. The estimators are compared to the method of composite estimation currently used in the CPS. The paper is organized as follows: In section 2, we state some basic assumptions regarding the general class of repeated surveys considered in this paper. A description of the CPS method of composite estimation is given in section 3. The method of best linear unbiased estimation is discussed in section 4. In section 5, we present a recursive regression estimation procedure designed to reduce the computational complexity associated with best linear unbiased estimation. Section 6 is devoted to an application to data from the CPS. Alternative estimators and rotation designs are compared.

2. BASIC ASSUMPTIONS

In this section, we describe surveys of the type we will study. A rotation group is a set of individuals selected for the sample and observed for a fixed number of periods and in a fixed pattern over time. Assume that in each period of the survey, s rotation groups are included in the sample, where $s > 1$ is fixed. Assume that the basic data from the survey can be organized in a set of elementary estimators (such as simple sample means and estimated totals) of the parameters of interest (such as population means and totals), where a set of elementary estimators is associated with each rotation group. For computational convenience, the data for p periods can be arranged in a $p \times s$ data matrix, denoted by $\boldsymbol{H}$, in such a way that the observations on a rotation group appear in only one column. The total

[1] Ibrahim S. Yansaneh, Statistical Group, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850; and Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50011 U.S.A.

number of elementary estimators is $n = p \times s$. We call the columns of $\boldsymbol{H}$ streams. Several rotation groups can appear in a stream. Assume that:

(1) A given rotation group in a stream is observed over a period of total length $m + 1$, and the observation pattern for rotation groups is fixed and is the same for all groups.

(2) The design is balanced on time-in-sample. That is, of the s rotation groups included in the sample at a given time, one group is being observed for the first time, one is being observed for the second time, ..., one is being observed for the last time, where the last time is separated by m periods from the first observation.

These assumptions are satisfied by surveys such as the CPS and the Canadian Labor Force Survey. See Yansaneh (1997) for an illustration of the 4-8-4 rotation scheme used in the CPS.

3. THE CPS COMPOSITE ESTIMATOR

In general, composite estimators combine recent estimator(s) and data from the current and preceding period(s) to form an estimator for the current period. With the CPS, six of the eight rotation groups observed at time t were observed at time $t - 1$. We shall refer to these six rotation groups as continuing rotation groups, and the remaining two as incoming rotation groups.

The composite estimator currently in use is determined by two parameters. The estimator is

$$\hat{\theta}_{t,c} = (1 - \pi_1)\bar{y}_t + \pi_1(\hat{\theta}_{t-1,c} + \hat{\delta}_{t,t-1}) + \pi_2\hat{\delta}_t \qquad (1)$$

where, for the estimator currently used, $\pi_1 = 0.4$ and $\pi_2 = 0.2$, $y_{t,k}$ is the elementary estimate of level obtained from the rotation group which is in its k-th time in sample at time t, $\bar{y}_t = 8^{-1}\sum_{k=1}^{8} y_{t,k}$ is the basic estimator, defined as the mean of the elementary estimates based on the eight rotation groups observed at time t, $\hat{\theta}_{t-1,c}$ is the composite estimator for time $t - 1$, $\hat{\delta}_{t,t-1}$ is an estimate of change in level, based on the six continuing rotation groups at time t, and $\hat{\delta}_t$ is the difference between the averages of the two incoming rotation groups and the six continuing rotation groups. Thus,

$$\hat{\delta}_{t,t-1} = 6^{-1}\sum_{k\in S}\left(y_{t,k} - y_{t-1,k-1}\right),$$

and

$$\hat{\delta}_t = 8^{-1}\left(\sum_{k\in T} y_{t,k} - 3^{-1}\sum_{k\in S} y_{t,k}\right),$$

where $T = \{1, 5\}$ and $S = \{2, 3, 4, 6, 7, 8\}$. The composite estimator used until 1985 contained only the first two terms on the right of (1). The third term was introduced for the purpose of reducing the time-in-sample effects appearing in the original estimator. The incoming rotation groups produce larger estimates of unemployed than do the continuing rotation groups. Therefore, the direct difference $\hat{\delta}_{t,t-1}$ is influenced by the fact that the rotation group in its first time-in-sample has a larger expected value than that of the second time-in-sample. The time-in-sample effects do not cancel out in the difference estimate. The third term is an adjustment term which has the effect of reducing both the variance of the original composite estimator and the bias associated with time-in-sample effects. See Bailar (1975) or Breau and Ernst (1983) for a discussion of the bias of the pre-1985 composite estimator due to time-in-sample effects. We shall refer to the three-term composite estimator currently used in the CPS as the CPS Composite Estimator. This estimator has a variance close to that of the best linear unbiased estimator for monthly estimates of unemployment level. Let $y_{i,t}$, $i = 1, 2, ..., s$, be the elementary estimator of the parameter of interest obtained from the rotation group which is in stream i at time t. The CPS composite estimator can be written as

$$\hat{\theta}_{t,c} = \sum_{i=1}^{8} \omega_{1,k(i,t)} y_{i,t} + \sum_{i=1}^{8} \omega_{2,k(i,t)} y_{i,t-1} + \pi_1\hat{\theta}_{t-1,c} \qquad (2)$$

where $k(i, t) = k$ defines the time-in-sample of observation (it) as a function of the stream (i) and time (t). If $\lambda_1 = 1/8$ and $\lambda_2 = -1/6$, and $\lambda_3 = 1/3$, then $\omega_{2,k} = \pi_1\lambda_2$, and

$$\omega_{1,k} = \begin{cases} (1 - \pi_1)\lambda_2 - \pi_1\lambda_2 - \pi_2\lambda_1\lambda_3 & \text{for} \quad k \in S \\ \lambda_1(1 - \pi_1 + \pi_2) & \text{for} \quad k \in T \end{cases}$$

Let

$$\boldsymbol{p}_1 = \left(\omega_{1,k(1,t)}, \omega_{1,k(2,t)}, \dots, \omega_{1,k(8,t)}\right)',$$

$$\boldsymbol{p}_2 = \left(\omega_{2,k(1,t)}, \omega_{2,k(2,t)}, \dots, \omega_{2,k(8,t)}\right)',$$

and $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{8,t})'$. Then,

$$\hat{\theta}_{t,c} = \boldsymbol{p}_i'\boldsymbol{y}_t + \boldsymbol{p}_2'\boldsymbol{y}_{t-1} + \pi_1\hat{\theta}_{t-1,c} \qquad (3)$$

Substituting in (3) recursively, we have, for an estimator initiated at time zero,

$$\hat{\theta}_{t,c} = \boldsymbol{p}_1'\boldsymbol{y}_t + \sum_{j=1}^{t} \pi_1^{j-1}(\boldsymbol{p}_2 + \pi_1\boldsymbol{p}_1)'\boldsymbol{y}_{t-1} \qquad (4)$$

Equation (4) is an expression of $\hat{\theta}_{t,c}$ as a linear function of current and past observations, where the weight of an observation declines as its distance from the current period increases.

4. BEST LINEAR UNBIASED ESTIMATION

Suppose $\Theta_p = (\theta_1, \theta_2, ..., \theta_p)'$ is the $p \times 1$ vector of parameters of interest, where θ_t, $t = 1, 2, ..., p$, is the level of the parameter of interest at time t. Thus at time j, θ_j is the current level of the parameter of interest. For example, in the context of the CPS, θ_j might represent the population mean or proportion of unemployed at time j. Our objective is to construct efficient estimators of the current level of the parameters. The change in level and average level over multiple periods of time are also of interest.

The best linear unbiased estimator (BLUE) of the current level is defined to be the minimum-variance unbiased linear combination of the elementary estimators from the rotation groups available for estimation. It is possible in the process of computing the BLUE for the current level, to also compute the BLUEs for all periods using data available at the current time.

Suppose that a repeated survey has been in operation for p periods and that s streams of data collected over p periods are available for estimation. Let $y_i = (y_{i,1}, y_{i,2}, ..., y_{i,p})'$ be the vector of p observations in the i-th stream at time t. Let Y_p be the data vector formed by the streams or columns of the $p \times s$ data matrix H, arranged chronologically. Thus, $Y_p = (y_1', y_2', ..., y_s')'$ is an $n \times 1$ vector of observations, where $n = s \times p$. Let $X = J_{s\times 1} \otimes I_{p\times p}$ be the $n \times p$ design matrix which relates the estimates in Y_p to their expected values in Θ_p; where $J_{s\times 1}$ is the $s \times 1$ vector of ones, $I_{p\times p}$ is the identity matrix of order p, and $\otimes$ denotes the Kronecker product. The linear model for Y_p is

$$Y_p = X\Theta_p + U_p \tag{5}$$

where U_p is the vector of error terms satisfying the assumptions $\mathrm{E}(U_p) = 0$ and $\mathrm{E}(U_p U_p') = V_p$, where V_p is assumed to be a known, symmetric, and positive definite matrix. By the Gauss-Markov Theorem, the BLUE of Θ_p is

$$\hat{\Theta}_p = (X' V_p^{-1} X)^{-1} X' V_p^{-1} Y_p.$$

The covariance matrix of $\hat{\Theta}_p$ is $\sum_p = (X' V_p^{-1} X)^{-1}$.

5. RECURSIVE REGRESSION ESTIMATION

Recursive estimation techniques have been found useful in situations where data do not all become available at the same time but rather accumulate over time, and the computation of optimal estimates based on all available data is impractical. See, for example, Odell and Lewis (1971), Sallas and Harville (1981) and references cited therein, for recursive algorithms for best linear unbiased estimation. Tiller (1989) presented a Kalman-filter approach to estimation of labor force characteristics using survey data.

As described in Section 4, the direct computation of the BLUE becomes progressively more complicated as the number of periods increases. We develop a recursive regression estimation procedure for repeated surveys that uses a judiciously chosen set of initial estimates, new observations of the current level, and the previous observations on the currently observed rotation groups to produce the BLUE of current level.

5.1 Transformed Elementary Estimates and a Proposed Estimator

Suppose a survey has been in operation for at least m periods and assume:

(3) The rotation groups are independent.

(4) The covariance structure of the observations is known.

(5) The covariance structure of the observations in a stream is constant over time, and it is the same for all streams.

These assumptions are used in the construction of a linear estimator. Assumption (3) will be relaxed for the computation of the variance of the estimator. Under assumptions (1) and (3), observations that are more than m periods apart are independent. At the current time, denoted by c, where $c > m$, a set of s elementary estimates of the parameter θ_c are observed. To construct the generalized least squares estimator, the s current observations are transformed so that they are uncorrelated with previous observations. After transformation, the expected values of the transformed observations are functions of θ_c and the parameters for the m preceding periods. Assume that the BLUE of the vector of parameters for the previous m periods, and the $m \times m$ covariance matrix of these estimators, are available. Thus, at time c, we have: (i) m initial estimates $\hat{\Theta}_{c-1(m)} = (\hat{\theta}_{c-m}, ..., \hat{\theta}_{c-1})'$; (ii) the covariance matrix $\sum_{c-1(m)}$ of $\hat{\Theta}_{c-1(m)}$; and (iii) s independent observations on the s streams at the current time. Let the transformed observations, denoted by z_{ic}, $i = 1, 2, ..., s$, be

$$z_{ic} = y_{i,c} - \sum_{j=1}^{m} b_{k(i,c),j}\, y_{i,c-j} \tag{6}$$

where $b_{k(i,c),j}$ are the coefficients such that $z_{i,c}$ is uncorrelated with $y_{i,c-j}$ for all $j > 0$. By assumptions (4) and (5), the coefficients $b_{k(i,c),j}$ are fixed over time. By assumption (3), $z_{i,c}$ is uncorrelated with all earlier observations. The expected value of $z_{i,c}$ is $\theta_c - \sum_{j=1}^{m} b_{k(i,c),j}\theta_{c-j}$, $i = 1, 2, ..., s$.

5.2 The Recursive Regression Estimator

Let $\hat{\theta}_h(t)$, $h \le t$, denote the least squares estimator of the (scalar) parameter θ_h constructed using data through time t; and let $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), ..., \hat{\theta}_t(t))'$ denote the least squares estimator of the vector of m parameters $\theta_{t-m+1}, ..., \theta_t$, at time t constructed using data through time t. Our objective is to construct the minimum variance estimator for θ_c, the current level of the parameter of interest using all data available at time c. A linear model for data available at the current time is

$$Z_c = W\Theta_{c(m+1)} + U_c \tag{7}$$

where

$$W = \begin{pmatrix} I_m & 0 \\ X_{21} & J_s \end{pmatrix},$$

$Z_c' = (\hat{\Theta}_{c-1(m)}', z_c'), z_c' = (z_{1c}, \ldots, z_{sc})$, and X_{21} is an $s \times m$ matrix whose entries are constant over time, and are functions of the coefficients $b_{k,j}$ of (6). If $\mathrm{Var}\{z_{i,c}\} = \sigma_i^2, i = 1, 2, \ldots, s$, and Ω_{22} is the diagonal matrix with σ_i^2 as the diagonal entries, then the covariance matrix of Z_c is $V_c = \mathrm{blockdiag}\{\sum_{c-1(m)}, \Omega_{22}\}$. It is assumed that σ_i^2, $i = 1, 2, \ldots, s$, are positive.

The recursive regression estimator (RRE) of $\Theta_{c(m+1)}$ is defined to be the least squares estimator of $\Theta_{c(m+1)}$ based on model (7). Thus the RRE of $\Theta_{c(m+1)}$ is

$$\hat{\Theta}_{c(m+1)} = (W' V_c^{-1} W)^{-1} W' V_c^{-1} Z_c \tag{8}$$

and the covariance matrix of $\hat{\Theta}_{c(m+1)}$ is $Q_{c(m+1)} = (W' V_c^{-1} W)^{-1}$.

The utility of the estimator (8) is its computational simplicity. At any fixed time t in a repeated survey, all the information relevant to the problem of estimating $\theta_t, \theta_{t-1}, \ldots, \theta_{t-m}$ can be obtained from a set of m recursive least squares estimates and the current observations.

We now describe more fully the recursive regression procedure. At time t, we have $\hat{\Theta}_{t(m+1)}$, the RRE of $\Theta_{t(m+1)}$, and its $(m+1) \times (m+1)$ covariance matrix $\sum_{t(m+1)}$. Partition $\sum_{t(m+1)}$ as

$$\sum_{t(m+1)} = \begin{pmatrix} v_{11,t} & V_{12,t} \\ V_{12,t}' & \sum_{t(m)} \end{pmatrix},$$

where $v_{11,t}$ is the variance of $\hat{\theta}_{t-m}(t), \sum_{t(m)}$ is the covariance matrix of $(\hat{\theta}_{t-m+1}(t), \ldots, \hat{\theta}_t(t))'$, and $V_{12,t}$ is the covariance between these two quantities. Observe that if θ_{t-m} is retained in the parameter vector and $\hat{\theta}_{t-m}$ is retained in the data vector, the estimator of θ_{t+1} is unchanged (the estimator of θ_{t-m} would, in general, be changed). This is because the estimator of the original parameter vector of a least squares problem is not changed if an observation whose expectation is equal to a single new parameter is added to the problem. Thus, to update the RRE for the next period, we drop the initial estimate for the earliest period, $\hat{\theta}_{t-m}(t)$, from the data vector, and drop the corresponding parameter θ_{t-m} from the parameter vector. The parameter θ_{t+1} is then added to the parameter vector. In this way, the dimension of the basic model matrix W of the estimation problem is kept constant over time. Thus in the class of repeated surveys considered in this paper, there is an upper bound on the computational effort required for the BLUE of the vector of parameters of interest.

The model at time $t+1$ may be written as model (7), with $c = t+1, Z_{t+1} = (\hat{\theta}_{t-m+1}(t), \ldots, \hat{\theta}_{t-1}(t), \hat{\theta}_t(t), z_{t+1}')'$, $\Theta_{t+1(m+1)} = (\theta_{t-m+1}, \ldots, \theta_t, \theta_{t+1})'$, and the covariance matrix of Z_{t+1} is $V_{t+1} = \mathrm{blockdiag}\{\sum_{t(m)}, \Omega_{22}\}$. The BLUE of $\Theta_{t+1(m+1)}$ and its covariance matrix are then obtained from the usual least squares formulas. The least squares estimators of the last m elements of $\Theta_{t+1(m+1)}$ are then used as the initial estimates in the model for the next iteration.

The following theorem states that the covariance matrix of the vector of recursive least squares estimators converges to a positive definite matrix as the number of periods in the survey increases indefinitely. A proof is given in the appendix.

Theorem: At any time t, let the vector of recursive least squares estimators $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), \ldots, \hat{\theta}_{t-1}(t), \hat{\theta}_t(t))'$ be the BLUE of the vector of parameters $\Theta_{t(m)} = (\theta_{t-m+1}, \ldots, \theta_{t-1}, \theta_t)'$ based on data through time t. Let $\sum_{t(m)}$ be the covariance matrix of $\hat{\Theta}_{t(m)}$. Let the assumptions (1) through (5) hold. Also assume that the elements of V_n^{-1} are bounded for all n, where V_n is the covariance matrix of any n observations. Then, the covariance matrix $\sum_{t(m)}$ converges as $t \to \infty$; and the limit is an $m \times m$ positive definite matrix.

6. APPLICATION TO THE U.S. CURRENT POPULATION SURVEY

6.1 The CPS Design

The CPS is a monthly household survey conducted by the United States Census Bureau in cooperation with the Bureau of Labor Statistics for the purpose of providing national estimates of labor force characteristics such as the number employed, unemployed, and in the civilian labor force; and other characteristics of the non-institutionalized civilian population. The sample design of the CPS contains a rotation scheme that includes the replacement of a fraction of the households in the sample each month. For any given month, the sample consists of eight time-in-sample panels or rotation groups, of which one is being interviewed for the first time, one is being interviewed for the second time,..., and one is being interviewed for the eighth time. In other words, the interview scheme is balanced on time-in-sample. Households in a rotation group are interviewed for four consecutive months, dropped for the next eight succeeding months, and then interviewed for another four consecutive months. They are then dropped from the sample entirely. This system of interviewing is called the 4-8-4 rotation scheme, and is a special case of schemes described by Rao and Graham (1964).

6.2 Estimation and Variance Estimation Procedures

We use estimates of the covariance structure of data from the CPS to compare alternative estimators and rotation designs. See Adam and Fuller (1992) and Fuller, Adam and Yansaneh (1993) for a detailed description of the construction of the model, the estimation of its parameters, and the estimation of the covariance structure of observations within a given rotation group for various characteristics of interest. Because the rotation groups come from the same set of primary sampling units, they are not independent and a component is included in the covariances to reflect the fact that the primary sampling units do not change. The RRE is computed with the eight current simple estimators and the 15 estimators for the 15 preceding periods. In computing the RRE, the covariances are used to create eight linear combinations of the current and the preceding fifteen observations that are uncorrelated with the preceding fifteen observations. Because of the primary sampling unit effect, these linear combinations are correlated with observations more than 15 periods in the past and in the same stream. Hence, they are correlated with the preceding estimators. The correlations with earlier estimators, $\hat{\theta}_{t-i}$, $i = 1, ..., 15$, are included in the covariance matrix when the estimator of θ_t is constructed. However, because only the most recent 15 observations are used, the resultant estimator of θ_t is not the BLUE for current level. The calculated covariance matrix of $(\hat{\theta}_{t-15}, ..., \hat{\theta}_{t-1}, \hat{\theta}_t)'$ is correct and, because the primary sampling unit effect is modest, it is felt that the estimator has efficiency close to that of the BLUE.

We shall restrict attention to the estimation of various parameters for two characteristics of interest: Employed and Unemployed. For each characteristic, the parameters of interest are the current level and period-to-period change for up to 12 periods. The estimators considered for comparison are the CPS composite estimator; the RRE; and the BLUEs using 2, 3, 12, 16, and 24 periods, where the BLUE for p periods at time t is the least squares estimator constructed using data from time $t - p + 1$ through time t. Results are reported for BLUEs based on 12 and 16 periods. In following the practice of the U.S. Bureau of Labor Statistics for CPS estimators, the estimators are not modified as new data become available. Thus the estimator of change in level of a characteristic of interest between times $t - 1$ and t is not the best possible estimator given all available data. It is the difference between the best estimator at time t based on data through time t and the best estimator at time $t - 1$ based on data through time $t - 1$.

We do not consider seasonal adjustment in this discussion. However, the estimation procedures presented can be extended to include seasonal adjustments. To compute the variance of a given estimator at a given time, the estimator is first expressed as a linear combination of all the observations available at that time. The variance of the estimator is then computed as a function of the coefficients of the linear combination and the entries of the covariance matrix.

6.3 Numerical Results and Discussion

6.3.1 Comparison of Alternative Estimators

The variances of the alternative estimators relative to the variance of the basic estimator of current level, for each of the characteristics of interest, are presented in Table 1. Recall that the basic estimator of the current level, denoted by $\bar{y}_t$ is the simple mean of the eight elementary estimators obtained from the eight rotation groups observed at time t. That is, $\bar{y}_t = 8^{-1} \sum_{k=1}^{8} y_{t,k}$, and $\mathrm{Var}(\bar{y}_t) = \sigma^2/8$, where $\sigma^2 = \mathrm{Var}(y_{t,k})$ for all t and k. The basic estimator of change between two periods is the difference between the simple means for the two periods.

The BLUE procedure based on 3 periods or more produces more efficient estimators of current level than the CPS composite estimator. In general, the best linear unbiased estimation procedure becomes more statistically efficient as the number of periods increases. For both characteristics, the results reveal that the best linear unbiased procedure based on 12 periods is uniformly more efficient than the CPS composite estimator for all parameters, except one-period change in unemployed. Recall that the estimator of change is not BLUE because the estimator is the difference of estimators constructed at time t and at time $t - 1$. Thus, the estimator called "BLUE" is best only for current level using the stated amount of data. The difference between the variance of the composite estimator of one-period change and the variance of the 12-period BLUE of one-period change in unemployed is less than one percent. The gain in precision of the best linear unbiased estimation procedure for employed relative to the CPS composite estimator for current level is 22% for the BLUE for 12 periods, 28% for the BLUE for 16 periods, 30% for the BLUE for 24 periods, and 33% for the RRE. The corresponding gains for unemployed are 2%, 3%, and 3%. These results are a reflection of the nature of the autocorrelation functions of the characteristics. The autocorrelation function for unemployed declines much faster than that for employed.

With the exception of one-period change in employed, there is an improvement in the efficiency of the estimation of change from using the alternative estimators instead of the CPS composite estimator. The gain in precision increases as the number of periods increases, reaching a maximum value at five-period change for both characteristics. The gain then decreases slightly. In the case of the RRE, the maximum gain in efficiency for estimated change is 64% for employed and 5% for unemployed.

Table 1
Variances of alternative estimators relative to the variance of the basic estimator of current level

	Employed				Unemployed			
Parameter	CPS Comp.	BLUE for 12 periods	BLUE for 16 periods	Recursive Regression Estimator	CPS Comp.	BLUE for 12 Periods	BLUE for 16 periods	Recursive Regression Estimator
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.704	0.672	0.650	0.947	0.924	0.918	0.918
1-period change	0.511	0.457	0.437	0.432	1.070	1.077	1.073	1.073
2-period change	0.813	0.646	0.613	0.604	1.361	1.345	1.338	1.338
3-period change	1.065	0.763	0.724	0.711	1.528	1.481	1.473	1.473
4-period change	1.279	0.830	0.800	0.784	1.645	1.569	1.563	1.562
5-period change	1.363	0.880	0.847	0.829	1.691	1.614	1.607	1.606
6-period change	1.390	0.910	0.873	0.855	1.708	1.637	1.628	1.628
7-period change	1.388	0.930	0.884	0.865	1.710	1.646	1.637	1.636
8-period change	1.353	0.932	0.884	0.860	1.701	1.645	1.635	1.634
9-period change	1.255	0.912	0.854	0.832	1.671	1.624	1.614	1.614
10-period change	1.154	0.895	0.824	0.806	1.641	1.606	1.595	1.595
11-period change	1.061	0.883	0.795	0.782	1.614	1.590	1.578	1.578
12-period change	0.992	0.883	0.767	0.761	1.593	1.577	1.563	1.563

6.3.2 Comparison of Alternative Estimators and Rotation Designs

The variances of alternative estimators under various rotation designs are given in Table 2. All variances are relative to the variance of the basic estimator of current level under that design. The efficiencies of alternative estimators of current level, change in level, and average level for multiple time periods are compared under the intermittent 4-8-4 rotation design and two continuous rotation designs. The continuous rotation designs are the 6-continuous scheme and the 8-continuous scheme. The 6-continuous scheme is the rotation scheme used in the Canadian Labor Force Survey conducted by Statistics Canada. For each period of the survey, the sample consists of six rotation groups, one rotation group in its first time-in-sample, ..., and one rotation group in its sixth time-in-sample. A given rotation group remains in the sample for six consecutive periods and then permanently drops out of the sample. See Kumar and Lee (1983) for more details about the design of the Canadian Labor Force Survey. In the 8-continuous scheme, there are 8 rotation groups in the sample for each period. A given rotation group remains in the sample for eight consecutive periods and then permanently drops out of the sample.

We compare the performance under the various rotation designs using the BLUE of current level based on 36 periods. We call this estimator the "best estimator" because its efficiency is vitually the same as that of the RRE. For all rotation schemes under consideration, there are some improvements in the precision of the estimators of current level from using the best estimator relative to the CPS composite estimator. As seen in Table 2, the gain is highest for employed where, under the 4-8-4 rotation scheme, the variance of the best estimator of current level is only 76% of that of the CPS composite estimator.

The precision of the estimators of change relative to the precision of the CPS composite estimator depends on the rotation design. From Table 2, we see that under the 4-8-4 rotation scheme, there is some gain in precision, which increases as the lag increases. For employed, the variance of the least squares estimator is 85% of the variance of the CPS composite estimator for one-period change, 61% of the variance of the CPS composite estimator for six-period change, and 76% of the variance of the CPS composite estimator for 12-period change. (Compare columns (2) and (3) of Table 2.)

Table 2
Variances of alternative estimators and rotation designs; the variance of the basic estimator of current level under each design equals one

	Employed				Unemployed			
Parameter	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.653	0.761	0.759	0.947	0.918	0.944	0.938
1-period change	0.511	0.432	0.395	0.434	1.070	1.073	1.003	1.051
2-period change	0.813	0.604	0.559	0.619	1.361	1.338	1.250	1.312
3-period change	1.065	0.710	0.669	0.747	1.528	1.473	1.372	1.443
4-period change	1.279	0.783	0.731	0.829	1.645	1.562	1.473	1.543
5-period change	1.363	0.828	0.782	0.901	1.691	1.606	1.533	1.607
6-period change	1.390	0.854	0.828	0.970	1.708	1.628	1.577	1.655
7-period change	1.388	0.863	0.874	1.026	1.710	1.636	1.612	1.686
8-period change	1.353	0.858	0.828	0.960	1.701	1.934	1.642	1.705
9-period change	1.255	0.830	0.960	1.108	1.671	1.614	1.663	1.719
10-period change	1.154	0.803	0.993	1.139	1.641	1.595	1.678	1.727
11-period change	1.061	0.779	1.021	1.165	1.614	1.578	1.688	1.733
12-period change	0.992	0.758	1.046	1.186	1.593	1.564	1.696	1.737
12-period average	0.369	0.326	0.440	0.394	0.255	0.249	0.301	0.266
12-change in averages	0.248	0.162	0.365	0.403	0.273	0.262	0.372	0.359

For estimating 12-period averages in employed using the 4-8-4 design, the CPS composite estimator is about 13% less efficient than the least squares estimator and, for estimating change in 12-period averages, it is about 53% less efficient, as can be seen by comparing the second and third columns of Table 2. For unemployed and the 4-8-4 design, there are only modest gains in precision from using the least squares estimator relative to the CPS composite estimator, as shown in the sixth and seventh columns of Table 2.

For estimation of 12-period change, 12-period average and change in 12-period averages, the 4-8-4 design is much superior to both continuous rotation designs for both characteristics. The continuous designs are generally superior for period-to-period changes for short periods.

6.3.3 Internal Consistency

In our analysis, we have constructed the best estimator of employed using only the past history of employed and the best estimator of unemployed using only the past history of unemployed. There is no formal reason not to include the past history of both employed and unemployed in the construction of the estimators. However, Fuller *et al.* (1993) state that the estimated cross correlations are less than 0.10, suggesting that there is little gain from such inclusion.

A method of constructing estimates of multiple characteristics that are internally consistent was suggested by Fuller (1990). In this procedure, estimates of employed, unemployed, and not-in-the-labor-force are constructed. Then these estimates are used as controls in a regression procedure to construct weights for the current observations. The weights can then be used to construct internally consistent estimates of any parameter of interest. The estimation procedure, including estimates of subdivisions of the labor force, is planned for implementation in 1998 for the CPS. See Lent, Miller and Cantwell (1996).

6.4 Conclusions

The main conclusions emerging form the variance computations in this section can be summarized as follows:

1. For all rotation designs and all characteristics under consideration, there are alternative estimation procedures with a variance of the current level smaller than that of the CPS composite estimator.
2. For estimation of change under the 4-8-4 rotation design, the gain in precision of the alternative estimators relative to the CPS composite estimator increases as the lag increases, and peaks around the lag of minimum overlap.
3. The intermittent 4-8-4 rotation design is inferior to the continuous rotation designs for short-period changes, but superior for current level, long-period averages, and changes in long-period averages.
4. The CPS composite estimator is comparable to the RRE for unemployed for the estimation of one-period change and 12-period change. However, the recursive regression estimation procedure is superior to the CPS composite estimator for other measures of change.
5. The RRE is more efficient in estimating change in level at lags for which the CPS composite estimator is not targeted, for instance, lags of four months to six months.

ACKNOWLEDGMENTS

The authors thank the referee and John Eltinge for helpful comments on earlier versions of this paper. Research for this paper was partly supported by Joint Statistical Agreement 91-21 with the U.S. Bureau of the Census and Cooperative Agreement 43-3AEU-80088 with the National Agricultural Statistics Service and the U.S. Bureau of the Census. The views expressed in this paper are those of the authors and do no necessarily represent the policies of the Bureau of the Census, the Bureau of Labor Statistics, or the National Agricultural Statistics Service.

APPENDIX

Lemma 1. Let the assumptions of the theorem hold. Then the variance of the estimator of current level θ_c converges to a positive number as the number of periods increases.

Proof. If the means $\theta_{c-1}, \theta_{c-2}, ..., \theta_{c-m}$, were known, then $g_{1c}, i = 1, 2, ..., s$ are unbiased estimators of θ_c, where $g_{1c} = y_{1,c}; g_{2,c} = y_{2,c} - b_{21}(y_{2,c-1} - \theta_{c-1}); ...;$ and $g_{sc} = y_{sc} - \sum_{j=1}^{m} b_{sj}(y_{s,c-j} - \theta_{c-j})$. Furthermore, $g_{1c}, i = 1, 2, ..., s$ are independent with variances $\sigma_i^2, i = 1, 2, ..., s$. We may write the linear model:

$$\boldsymbol{g} = \boldsymbol{J}_s\theta_c + e \tag{A1}$$

where $\boldsymbol{g} = (g_{1c}, g_{2c}, ..., g_{sc})'$, $\boldsymbol{J}_s$ is the $s \times 1$ column vector of ones, and $\boldsymbol{e}$ is the $s \times 1$ vector of errors with $E(\boldsymbol{e}) = 0$, and $E(\boldsymbol{ee}') = \boldsymbol{V}_s = \text{Diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_s^2\}$. Thus the BLUE of θ_c for model (A1) has variance $(\sum_{i=1}^{s} \sigma_i^{-2})^{-1}$. By assumption, the variances $\sigma_i^2, i = 1, 2, ..., s$ are bounded below and the quantity $(\sum_{i=1}^{s} \sigma_i^{-2})^{-1}$ is a positive lower bound for the variance of the estimator of θ_c [see Lemma 4.2.3 of Yansaneh (1992)]. The variance of the estimator of θ_c is non-increasing as the number of observations increases, and hence, the variance converges to a positive number.

Lemma 2. Let the assumptions of the theorem hold. Then the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1}, ..., \theta_{t-1}$, based on data through time t, converges to a positive number as t increases.

Proof. First, suppose at a fixed time τ, at least m periods of observations are available both prior to τ and after τ. Define a transformation of the following form for the observations in each of the s streams at time τ: $u_{i\tau} = y_{i,\tau} - \sum_{j=-m}^{m} b_{k(i,\tau),j} y_{i,\tau-j}$, where $b_{k(i,\tau),0} = 0$ and $u_{i\tau}$ is uncorrelated with all observations preceding and succeeding $y_{i,\tau}$ in the i-th stream. Let the variance of $u_{i\tau}$ be λ_i^2, $i = 1, 2, ..., s$. These variances are bounded below by assumption. We conclude, as before, that there is a positive lower bound for the diagonal elements of the covariance matrix of the vector of recursive least squares estimators.

Now, assume that at time t, we begin the sequence of estimation with the vector of recursive least squares estimators $\hat{\Theta}_{t-1(m)} = (\hat{\theta}_{t-m}, ..., \hat{\theta}_{t-1})'$ based on data for the preceding m periods; and the vector of transformed observations $z_t' = (z_{1t}, ..., z_{st})$. Thus the linear model for the data at time t is given by (7), with c replaced by t. The data vector $\boldsymbol{Z}_t$ is of fixed dimension. Therefore, the covariance matrix of the BLUE of the vector of parameters $\Theta_{t(m+1)} = (\theta_{t-m}, ..., \theta_{t-1}, \theta_t)'$ is $\sum_{t(m+1)} = (\boldsymbol{W}' \boldsymbol{V}_t^{-1} \boldsymbol{W})^{-1}$. For computational convenience, we express $\boldsymbol{W}$ as $(\boldsymbol{I}_{m+1}, \boldsymbol{M}')'$, where $\boldsymbol{I}_{m+1}$ is the identity matrix of order $m + 1$, and $\boldsymbol{M}$ is an $(s - 1) \times (m + 1)$ matrix which is constant over time. Thus we have

$$\begin{aligned}\sum\nolimits_{t(m+1)} &= (\Omega_{t-1(m+1)}^{-1} + \boldsymbol{M}' \Omega_{00}^{-1} \boldsymbol{M})^{-1} \\ &= \Omega_{t-1(m+1)} - \Omega_{t-1(m+1)} \boldsymbol{M}' \boldsymbol{D}_t^{-1} \boldsymbol{M} \Omega_{t-1(m+1)}\end{aligned} \tag{A2}$$

where

$\Omega_{t-1(m+1)} = \text{blockdiag}\{\sum_{t-1(m)}, \sigma_1^2\}$, $\Omega_{00} = \text{diag}\{\sigma_2^2, ..., \sigma_s^2\}$, and $\boldsymbol{D}_t = \Omega_{00} + \boldsymbol{M}\Omega_{t-1(m+1)}\boldsymbol{M}'$. Since the second term on the right hand side of (A2) is positive definite, we conclude that the first m diagonal elements of $\sum_{t(m+1)}$ are less than or equal to the original diagonal elements of $\sum_{t-1(m)}$. This means that as t increases, the variances of the estimators of $\theta_{t-m}, ..., \theta_{t-2}, \theta_{t-1}$ are non-increasing. Since these variances are bounded below by a positive quantity, we conclude that the variances of the estimators of $\theta_{t-m}, ..., \theta_{t-2}, \theta_{t-1}$ converge to positive numbers as t increases.

Lemma 3. Let the assumptions of the theorem hold. Then, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, based on data through time t, converges to a positive number as t increases.

Proof. First, we show that variance of the least squares estimator of $\theta_c - \theta_{c-1}$ (where denotes the current period) converges as the number of periods increases by mimicking the arguments in the proof of Lemma 1. Also, arguments similar to those in the proof of Lemma 2 can be used to show that the variances of the least squares estimators of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, all converge as the number of periods increases.

Proof of theorem. Since $\sum_{t(m)}$ is a submatrix of the covariance matrix $\sum_{t(m+1)}$ of the least squares estimators of the full set of parameters $\theta_{t-m}, \theta_{t-m+1}, ..., \theta_{t-1}, \theta_t$, at time t, it is enough to show that $\sum_{t(m+1)}$ converges to a positive definite matrix as $t \to \infty$. From Lemma 1 and Lemma 2, each of the diagonal elements of $\sum_{t(m+1)}$ converges to a positive number as $t \to \infty$. From Lemma 3, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, converges to a positive number as $t \to \infty$. It follows that for each j, $1 \le j \le m$, the covariance between the least squares estimators of θ_t and θ_{t-j} converges as $t \to \infty$ and hence the covariance matrix $\sum_{t(m+1)}$ converges as $t \to \infty$.

Next, we prove that the limiting covariance matrix is positive definite. Let $\lim_{t \to \infty} \sum_{t(m)} = \sum_{(m)}$. It is enough to show that the variance of any non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 1, 2, ..., m$, is bounded below by a positive quantity. Let v_{mm} be the lower bound of every linear combination of the observations with one of the coefficients equal to one. The bound is positive by the assumption that the elements of V_n^{-1} are bounded.

Now, every estimator of the parameter θ_{t-j}, $j = 0, 1, ..., m$ is a linear combination of all observations such that the sum of the coefficients for the observations in the s streams at time $t - j$ is one, and the sum of the coefficients for the observations in the s streams at any other time is zero. This is a condition for the unbiasedness of the estimator for time $t - j$. For the sum of the coefficients of the s observations at time $t - j$ to be equal to one, at least one of the coefficients must be greater than or equal to s^{-1}. The minimum variance of any linear combination with first coefficient equal to s^{-1} is $s^{-2} v_{mm}$. Therefore, for $j = 0, 1, ..., m$, $\mathrm{Var}\{\hat{\theta}_{t-j}(t)\} \ge s^{-2} v_{mm}$.

Now, consider an arbitrary, non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 0, 1, ..., m$, given by $\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t)$, where, without loss of generality, $\gamma_0 = 1$. This linear combination can be expressed as

$$\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t) = \hat{\theta}_t(t) + \sum_{j=1}^{m} \gamma_j \hat{\theta}_{t-j}(t)$$

$$= \sum_{i=1}^{s} \sum_{h=1}^{t} c_{ih} y_{i,h} + \sum_{j=1}^{m} \gamma_j \sum_{i=1}^{s} \sum_{h=1}^{t} f_{ih(t-j)} y_{i,h} \tag{A3}$$

$$= \sum_{i=1}^{s} \left[c_{it} + \sum_{j=1}^{m} \gamma_j f_{it(t-j)} \right] y_{i,t} + \sum_{i=1}^{s} \sum_{h=1}^{t-1} \left[c_{ih} + \sum_{j=1}^{m} \gamma_j f_{ih(t-j)} \right] y_{i,h}$$

where c_{it}, $i = 1, 2 ..., s$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_t(t)$, and $f_{it(t-j)}$, $j = 1, ..., m$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_{t-j}(t)$, $j = 1, ..., m$, respectively. Therefore, $\sum_{i=1}^{s} c_{it} = 1$, and $\sum_{i=1}^{s} f_{it(t-j)} = 0$, for $j = 1, ..., m$. Thus $\sum_{i=1}^{s} [c_{it} + \sum_{j=1}^{m} \gamma_j f_{it(t-j)}] = 1$. That is, in the linear combination (A3), the sum of the coefficients for the observations $y_{i,t}$, $i = 1, 2, ..., s$, at time t is one. Therefore, at least one of the coefficients is greater than or equal to s^{-1}. Hence, $\mathrm{Var}\{\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t)\} \ge s^{-2} v_{mm}$, and we conclude that $\sum_{(m)}$ is positive definite.

REFERENCES

ADAM, A., and FULLER, W.A. (1992). Covariance estimators for the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 586-591.

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

BELL, W.R., and HILLMER, S.C. (1990). The time series approach to estimation for periodic surveys. *Survey Methodology*, 16, 195-215.

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

BREAU, P., and ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1993). Estimators for longitudinal surveys. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, Ottawa, Canada, 309-324.

JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

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

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 1-24.

LENT, J., MILLER, S.M., and CANTWELL, P.J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 130-139.

ODELL, P.L., and LEWIS, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.

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

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

RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

SCOTT, A.J., SMITH, T.M.F, and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

SINGH, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.

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

YANSANEH, I.S. (1992). Least Squares Estimation for Repeated Surveys. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, Iowa.

YANSANEH, I.S. (1997). Recursive regression estimation in the presence of time-in-sample effects. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 164-169.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 41-50
Statistics Canada

Estimation of Variance of General Regression Estimator: Higher Level Calibration Approach

SARJINDER SINGH, STEPHEN HORN and FRANK YU[1]

ABSTRACT

In the present investigation, the problem of estimation of variance of the general linear regression estimator has been considered. It has been shown that the efficiency of the low level calibration approach adopted by Särndal (1996) is less than or equal to that of a class of estimators proposed by Deng and Wu (1987). A higher level calibration approach has also been suggested. The efficiency of higher level calibration approach is shown to improve on the original approach. Several estimators are shown to be the special cases of this proposed higher level calibration approach. An idea to find a non – negative estimate of variance of the GREG has been suggested. Results have been extended to a stratified random sampling design. An empirical study has also been carried out to study the performance of the proposed strategies. The well known statistical package, GES, developed at Statistics Canada can further be improved to obtain better estimates of variance of GREG using the proposed higher level calibration approach under certain circumstances discussed in this paper.

KEY WORDS: Calibration; Estimation of variance; Auxiliary information; Ratio and regression type estimators; Model assisted approach.

1. INTRODUCTION

The statisticians are often interested in the precision of survey estimates. The most commonly used estimator of population total/mean is the generalized linear regression (GREG) estimator. Let us consider the simplest case of the GREG where information on only one auxiliary variable is available. Consider a population $\Omega = \{1, 2, ..., N\}$, from which a probability sample $s\,(s \subset \Omega)$ is drawn with a given sampling design, $p(.)$. The inclusion probabilities $\pi_i = Pr(i \in s)$ and $\pi_{ij} \in Pr(i \text{ and } j \in s)$ are assumed to be strictly positive and known. Let y_i be the value of the variable of interest, y, for the i-th population element, with which also is associated an auxiliary variable x_i. For the elements, $i \in s$, we observe (y_i, x_i). The population total of the auxiliary variable x, $X = \sum_{i=1}^{N} x_i$, is assumed to be accurately known. The objective is to estimate the population total $Y = \sum_{i=1}^{N} y_i$. Deville and Särndal (1992) used calibration on known population x-total to modify the basic sampling design weights, $d_i = 1/\pi_i$, that appear in the Horvitz-Thompson (1952) estimator

$$\hat{Y}_{\text{HT}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i. \tag{1.1}$$

A new estimator,

$$\hat{Y}_{\text{DS}} = \sum_{i=1}^{n} w_i y_i \tag{1.2}$$

was proposed by Deville and Särndal (1992), with weights w_i as close as possible in an average sense for a given metric to the d_i, while respecting the calibration equation

$$\sum_{i=1}^{n} w_i x_i = X. \tag{1.3}$$

A simple case considered by Deville and Särndal (1992) is the minimization of chi-square type distance function given by

$$\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i} \tag{1.4}$$

where q_i are suitably chosen weights. In most of the situations, the value of $q_i = 1$. The form of the estimator depends upon the choice of q_i. By minimizing (1.4) subject to calibration equation (1.3) we obtain weights

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^{n} d_i q_i x_i^2} \left(X - \sum_{i=1}^{n} d_i x_i \right). \tag{1.5}$$

Substitution of the value of w_i from (1.5) in (1.2) leads to the traditional regression estimator of total given by

$$\hat{Y}_{\text{DS}} = \sum_{i=1}^{n} d_i y_i + \frac{\sum_{i=1}^{n} d_i q_i x_i y_i}{\sum_{i=1}^{n} d_i q_i x_i^2} \left(X - \sum_{i=1}^{n} d_i x_i \right). \tag{1.6}$$

In this paper, the problem of estimation of variance of the regression estimator (1.6) has been considered at two different levels of calibration. The higher level calibration approach covers a greater variety of estimators than the low level calibration approach adopted by Särndal (1996).

[1] Sarjinder Singh, Research Officer, Stephen Horn, Senior Research Officer and Frank Yu, Director, Methodology Division, The Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT 2616, Australia.

Higher level calibration approach makes use of known total as well as known variance of the auxiliary character, whereas low level calibration utilizes only known total of auxiliary character.

The section 4 has been devoted to study the stratified sampling design. The original stratum weights are calibrated which results in combined regression and combined ratio estimators in stratified sampling. The estimators of variance of combined regression and combined ratio estimators proposed by Wu (1985) are shown to be the special cases of the low level calibration approach. The higher level calibration approach has been shown to apply to a broader variety of estimators.

2. ESTIMATOR OF VARIANCE OF THE GREG: THE LOW LEVEL CALIBRATION APPROACH

Following model assisted survey sampling approach of Särndal, Swensson and Wretman (1989, 1992), the Yates-Grundy (1953) form of estimator of variance of the estimator (1.6) is given by

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{DS}}\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\left(w_i e_i - w_j e_j\right)^2 \qquad (2.1)$$

where $D_{ij} = (\pi_i \pi_j - \pi_{ij})/\pi_{ij}$, $i \neq j$ and $e_i = y_i - \hat{\beta} x_i$ have their usual meanings. This estimator can easily be written as

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{DS}}\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\,(d_i e_i - d_j e_j)^2 +$$

$$\hat{\psi}_1\left(X - \sum_{i=1}^{n} d_i x_i\right) + \hat{\psi}_2\left(X - \sum_{i-1}^{n} d_i x_i\right)^2 \qquad (2.2)$$

where

$$\hat{\psi}_1 = \frac{1}{\sum_{i=1}^{n} d_i q_i x_i^2}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\left(d_i e_i - d_j e_j\right)\left(d_i q_i x_i e_i - d_j q_j x_j e_j\right) \qquad (2.3)$$

and

$$\hat{\psi}_2 = \frac{1}{2\left(\sum_{i=1}^{n} d_i q_i x_i^2\right)^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\left(d_i q_i x_i e_i - d_j q_j x_j e_j\right)^2 \qquad (2.4)$$

The estimator at (2.1) has been discussed by Särndal *et al.* (1989, 1992, 1996) on different occasions and covers a variety of estimators as discussed below:

For simplicity, let us consider simple random sampling and without replacement (SRSWOR) design *i.e.*, $\pi_i = \pi_j = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$. Then we have following cases:

Case 2.1: If $q_i = 1$, then (1.6) reduces to the usual regression estimator of total, $\hat{Y}_{\mathrm{GREG}}$ (say). Now if $w_i = d_i$ in (2.1), it reduces to

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{GREG}}\right) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2 \qquad (2.5)$$

where $f = n/N$ and $e_i = y_i - \hat{\beta} x_i$. Thus (2.5) denotes the usual estimator of variance of the regression estimator (1.6).

Case 2.2: If $q_i = 1/x_i$ then the estimator (1.6) reduces to the ratio estimator of total, $\hat{Y}_{\mathrm{RATIO}}$ (say). The estimator (2.1) reduces to an estimator of variance of the estimator $\hat{Y}_{\mathrm{RATIO}}$, given by

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{RATIO}}\right) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2\left\{\frac{X}{\hat{X}}\right\}^2 \qquad (2.6)$$

where

$$e_i = y_i - \left(\frac{\bar{y}}{\bar{x}}\right) x_i \text{ and } \hat{X} = \frac{N}{n}\sum_{i=1}^{n} x_i.$$

The estimator at (2.6) is a special case of a class of estimators of variance of the ratio estimator proposed by Wu (1982) as

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{W}}\right) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2\left\{\frac{X}{\hat{X}}\right\}^g \qquad (2.7)$$

for $g = 2$.

Case 2.3: If $q_i = 1$ and w_i is given by (1.5) then (2.2) and hence (2.1) becomes

$$\hat{V}_{\mathrm{YG}}\left(\hat{Y}_{\mathrm{GREG}}\right) =$$

$$\frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2 + \hat{\psi}_1\left(X - \hat{X}\right) + \hat{\psi}_2\left(X - \hat{X}\right)^2 \qquad (2.8)$$

where

$$\hat{\psi}_1 = \frac{(N-n)}{\left(\sum_{i=1}^{n} x_i^2\right) n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(e_i - e_j)(x_i e_i - x_j e_j) \qquad (2.9)$$

and

$$\hat{\psi}_2 = \frac{(N-n)}{2N(n-1)\left(\sum_{i=1}^{n} x_1^2\right)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i e_i - x_j e_j)^2. \qquad (2.10)$$

Deng and Wu (1987) have defined a general class of estimators of the variance of the regression estimator as

$$\hat{V}_{\text{YG}}\left(\hat{Y}_{\text{DW}}\right) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2 \left\{\frac{X}{\hat{X}}\right\}^g \tag{2.11}$$

where $e_i = y_i - \hat{\beta} x_i$. The linear form of the class of estimators (2.11) takes the form as

$$\hat{V}_{\text{YG}}\left(\hat{Y}_{\text{DW}}\right) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2$$

$$\left[1 + g\left(\frac{X}{\hat{X}} - 1\right) + \frac{g(g-1)}{2}\left(\frac{X}{\hat{X}} - 1\right)^2 + \dots\right] \tag{2.12}$$

which is again similar to (2.8). Thus the low level calibration approach considers estimators of variance of estimators of total *i.e.*, both ratio and regression methods of estimation. It is remarkable that there is no choice of q_i which reduces (1.6) to the product method of estimation considered by Cochran (1963). Hence the estimation of variance of product estimator has not been considered. To look at the efficiency of such estimators, we consider an analogue of the general class of estimators for estimating variance of GREG by following Srivastava (1971) as

$$\hat{V}_{\text{S}}\left(\hat{Y}_{\text{GREG}}\right) = \left(\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2\right) H\left(\frac{X}{\hat{X}}\right) \tag{2.13}$$

where $H(.)$ is a parametric function such that $H(1) = 1$ and satisfies certain regularity conditions. Following Srivastava (1971), it is easy to see that analogues of the general class of estimators (2.13) attain the minimum variance of the class of estimators proposed by Deng and Wu (1987) for regression estimator and Wu(1982) ratio estimator. We want to say here that if we will attach any function of the ratio $X/\hat{X}$ to the usual estimator of variance given by

$$\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2,$$

the asymptotic variance of the resultant estimator remains the same. In other words, the efficiency of the estimators of variance of regression estimator (GREG) of total obtained through low level calibration remains less than or equal to the class of estimators proposed by Wu (1982) and Deng and Wu (1987). The weights w_i used to construct estimator of variance of GREG at (2.1) were obtained while estimating the population total and hence named as low level calibration weights for variance estimation. The next section is devoted to the higher level calibration approach where variance of auxiliary character is known. Several new estimators are shown as special cases of the proposed higher level calibration approach.

3. IMPROVED ESTIMATOR OF VARIANCE OF THE GREG: THE HIGHER LEVEL CALIBRATION APPROACH

Here we apply the calibration approach to estimate the variance of GREG estimator at (1.6). The weights D_{ij} of Yates and Grundy (1953) for an estimator of variance given at (2.1) are calibrated such that the estimator of variance for the auxiliary variable has the exact variance. We consider an estimator of variance of GREG

$$\hat{V}_{\text{SS}}\left(\hat{Y}_{\text{GREG}}\right) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{ij} (w_i e_i - w_j e_j)^2 \tag{3.1}$$

where Ω_{ij} are the modified weights attached to the quadratic expression by Yates and Grundy (1953) form of estimator and are as close as possible in an average sense for a given measure to the D_{ij} with respect to the calibration equation

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{ij} (d_i x_i - d_j x_j)^2 = V_{\text{YG}}\left(\hat{X}_{\text{HT}}\right) \tag{3.2}$$

where

$$V_{\text{YG}}\left(\hat{X}_{\text{HT}}\right) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij})(d_i x_i - d_j x_j)^2$$

denotes the known variance of the estimator of the auxiliary total $X (= \sum_{i=1}^{N} x_i)$ given by $\hat{X}_{\text{HT}} = \sum_{i=1}^{n} d_i x_i$. To compute the right hand side of (3.2) we need either information on every unit of the auxiliary character in the population, or only $V_{\text{YG}}(\hat{X}_{\text{HT}})$ obtained from a past survey or pilot survey. The examples of a situation where information on every unit of the auxiliary character is known are the establishment turnover recorded from census or administrative records or Business Register (BR) or Australian Taxation Office (ATO). Known variance of the auxiliary character has also been used by Das and Tripathi (1978), Singh and Srivastava (1980), Srivastava and Jhajj (1980, 1981), Isaki (1983), Singh and Singh (1988), Swain and Mishra (1992), Shah and Patel (1996) and Garcia and Cebrian (1996). Singh, Mangat and Mahajan (1995) have reviewed classes of estimators of unknown population parameters making use of the known variance of an auxiliary character. The idea of adjusting D_{ij} weights has also been discussed by Fuller (1970) through a regression type estimation procedure. For simplicity we restrict ourselves to the two dimensional Chi-Square (CS) type distance, D, between two $n \times n$ grids formed by the weights Ω_{ij} and D_{ij} for $i, j = 1, 2, \dots, n$, given by

$$D = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\Omega_{ij} - D_{ij})^2}{D_{ij}Q_{ij}} \qquad (3.3)$$

In most of the situations $Q_{ij} = 1$ but other types of weights can also be used. We will show that the ratio type adjustment using known variance of auxiliary character is a special case for a particular choice of Q_{ij}. Minimization of (3.3) subject to (3.2) leads to modified optimal weights given by

$$\Omega_{ij} = D_{ij} + \frac{D_{ij}Q_{ij}(d_i x_i - d_j x_j)^2}{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}Q_{ij}(d_i x_i - d_j x_j)^4}$$

$$\left[V_{\text{YG}}(\hat{X}_{\text{HT}}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(d_i x_i - d_j x_j)^2\right] \qquad (3.4)$$

for the optimal choice of Lagrange Multiplier λ, given by

$$\lambda = \frac{V_{\text{YG}}(\hat{X}_{\text{HT}}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(d_i x_i - d_j x_j)^2}{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\ Q_{ij}(d_i x_i - d_j x_j)^4}. \qquad (3.5)$$

Its proof is given in the Appendix. Substitution of Ω_{ij} from (3.4) in (3.1) leads to the following regression type estimator,

$$\hat{V}_{\text{SS}}(\hat{Y}_{\text{GREG}}) =$$

$$\hat{V}_{\text{YG}}(\hat{Y}_{\text{DS}}) + \hat{B}_1\left[V_{\text{YG}}(\hat{X}_{\text{HT}}) - \hat{V}_{\text{YG}}(\hat{X}_{\text{HT}})\right] \qquad (3.6)$$

where

$$\hat{B}_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}Q_{ij}(d_i x_i - d_j x_j)^2(w_i e_i - w_j e_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}Q_{ij}(d_i x_i - d_j x_j)^4}$$

$$= \frac{\hat{\mu}_{22}}{\hat{\mu}_{04}} \text{ (say)} \qquad (3.7)$$

$\hat{V}_{\text{YG}}(\hat{X}_{\text{HT}}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(d_i x_i - d_j x_j)^2$ and $\hat{V}_{\text{YG}}(\hat{Y}_{\text{DS}})$ is given in (2.1). Regression coefficient $\hat{B}_1$ makes use of the known total X of the auxiliary variable and hence can be treated as an improved estimator of regression coefficient by following Singh and Singh (1988). Under the higher level calibration approach, we have the following cases:

Case 3.1: Under SRSWOR sampling design if $q_i = x_i^{-1}$ and $Q_{ij} = (d_i x_i - d_j x_j)^{-2}$ are the weights attached at low level and higher level calibration approach, respectively, then the proposed strategy reduces to

$$\hat{V}_{\text{SS}}(\hat{Y}_{\text{Ratio}}) =$$

$$\frac{N^2(1-f)}{n} \times \frac{1}{(n-1)}\sum_{i=1}^{n} e_i^2\left(\frac{X}{\hat{X}}\right)^2\left(\frac{S_x^2}{s_x^2}\right) \qquad (3.8)$$

where $s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of $S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$.

Case 3.2: If $q_i = 1$ and $Q_{ij} = 1\ \forall\, i$ & j, then we have

$$\hat{V}_{\text{YG}}(\hat{Y}_{\text{GREG}}) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2 + \hat{\psi}_1(X - \hat{X}) +$$

$$\hat{\psi}_2(X - \hat{X})^2 + \hat{\psi}_3(S_x^2 - s_x^2) \qquad (3.9)$$

where $\hat{\psi}_1$ and $\hat{\psi}_2$ are given by (2.9) and (2.10), respectively, and

$$\hat{\psi}_3 = \frac{N^2(1-f)}{n\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^4}$$

$$\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{(x_i - x_j)(e_i - e_j) + \frac{(X - \hat{X})(x_i - x_j)^2}{\sum_{i=1}^{n} x_i^2}\right\}^2\right] \qquad (3.10)$$

Without loss of generality, the estimators of variance of GREG given at (3.8) and (3.9) are neither members of a low level calibration approach nor of the class of estimators by Deng and Wu (1987). These estimators are members of the analogues of classes of estimators for estimating variance of GREG given by Srivastava and Jhajj (1981) as

$$\hat{V}_{\text{SJ}}(\hat{Y}_{\text{GREG}}) = \left(\frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n} e_i^2\right) H\left(\frac{X}{\hat{X}}, \frac{S_x^2}{s_x^2}\right) \qquad (3.11)$$

where $H(.,.)$ is a parametric function such that $H(1, 1) = 1$ and which satisfies certain regularity conditions defined by them. Following Srivastava and Jhajj (1981) and Deng and Wu (1987), it is a class room exercise to see that the class of estimators at (3.11) remains better than the class of estimators defined at (2.11) and hence (2.13).

A difficult issue in using (3.1) is how to get non-negative estimates of variance using calibration. The simplest way is to optimize the CS distance function (3.3) subject to calibration constraint (3.2) along with the conditions $\Omega_{ij} \geq 0\ \forall\, i, j = 1, 2, \ldots, n$. While it is difficult to develop a solution to this problem theoretically, well known quadratic programming techniques can yield useful numerical results. Straightforward extension to using other distance functions, as discussed by Deville and Särndal (1992) for instance, to the two dimensional problem due to the indeterminate nature of the D_{ij} weights is not possible. It is open to others to propose new distance functions which guarantee the non-negativity of the weights.

4. STRATIFIED SAMPLING DESIGN

Suppose the population consists of L strata with N_h units in the h-th stratum from which a simple random sample of size n_h is taken without replacement. The total population size $N = \sum_{h=1}^{L} N_h$ and sample size $n = \sum_{h=1}^{L} n_h$. Associated with the i-th unit of the h-th stratum there are two values y_{h_i} and x_{h_i} with $x_{h_i} > 0$ being the covariate. For the h-th stratum, let $W_h = N_h / N$ be the stratum weights, $f_h = n_h / N_h$ the sample fraction, $\bar{y}_h$, $\bar{x}_h$, $\bar{Y}_h$, $\bar{X}_h$ the y- and x- sample and population means respectively. Assume $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ is known. The purpose is to estimate $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$, possibly by incorporating the covariate information x. The usual estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h . \tag{4.1}$$

We are considering a new estimator, given by

$$\bar{y}^*_{St} = \sum_{h=1}^{L} W_h^* \bar{y}_h \tag{4.2}$$

with new weights W_h^*. The new weights W_h^* are chosen such that chi-square type distance, given by

$$\sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h q_h} \tag{4.3}$$

is minimum subject to the condition

$$\sum_{h=1}^{L} W_h^* \bar{x}_h = \bar{X}. \tag{4.4}$$

Minimization of (4.3) subject to calibration equation (4.4) leads to the combined regression type estimator given by

$$\bar{y}^*_{St} = \sum_{h=1}^{L} W_h \bar{y}_h + \frac{\sum_{h=1}^{L} W_h q_h \bar{x}_h \bar{y}_h}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2} \left[\bar{X} - \sum_{h=1}^{L} W_h \bar{x}_h\right] \tag{4.5}$$

for the optimum choice of weights given by

$$W_h^* = W_h + \frac{W_h q_h \bar{x}_h}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2} \left(\bar{X} - \sum_{h=1}^{L} W_h \bar{x}_h\right) \tag{4.6}$$

If $q_h = \bar{x}_h^{-1}$ then estimator (4.5) reduces to the well known combined ratio estimator in stratified sampling. The well known estimator of variance of combined regression estimator is given by

$$\hat{V}\left(\bar{y}^*_{St}\right) = \sum_{h=1}^{L} \frac{W_h^2 \left(1 - f_h\right)}{n_h} s^2_{\hat{e}h} \tag{4.7}$$

where

$$s^2_{\hat{e}h} = (n_h - 1)^{-1} \sum_{i=1}^{n_h} e^2_{hi}$$

is the h-th stratum sample variance and $\hat{e}_{hi} = y_{hi} - \bar{y}_h - b(x_{hi} - \bar{x}_h)$ and $b = \sum_{h=1}^{L} W_h q_h \bar{y}_h \bar{x}_h / \sum_{h=1}^{L} W_h q_h \bar{x}_h^2$ have their usual meaning. The lower level calibration approach yields an estimator of variance of the combined regression estimator as

$$\hat{V}_c\left(\bar{y}^*_{St}\right) = \sum_{h=1}^{L} \frac{D_h W_h^{*2}}{W_h^2} s^2_{\hat{e}h} \tag{4.8}$$

where

$$D_h = \frac{W_h^2 (1 - f_h)}{n_h}$$

and W_h^* is given by (4.6). If $q_h = \bar{x}_h^{-1}$ then (4.8) reduces to an estimator given by

$$\hat{V}\left(\bar{y}^*_{St}\right)_{\text{RATIO}} = \left(\frac{\bar{X}}{\bar{x}_{St}}\right)^2 \sum_{h=1}^{L} \frac{W_h^2 \left(1 - f_h\right)}{n_h} s^2_{\hat{e}h} \tag{4.9}$$

which is a special case of a class of estimators for estimating the variance of combined ratio estimator given by Wu (1985) as

$$\hat{V}\left(\bar{y}^*_{St}\right)_W = \left(\frac{\bar{X}}{\bar{x}_{St}}\right)^g \sum_{h=1}^{L} \frac{W_h^2 \left(1 - f_h\right)}{n_h} s^2_{\hat{e}h} \tag{4.10}$$

for $g = 2$. The properties of variance estimators of the combined ratio estimator are also studied by Saxena, Nigham and Shukla (1995). In higher level calibration, a new estimator is given by

$$\hat{V}_{St}\left(\hat{Y}_{\text{GREG}}\right) = \sum_{h=1}^{L} \frac{\Omega_h W_h^{*2}}{W_h^2} s^2_{\hat{e}h} \tag{4.11}$$

where Ω_h are suitably chosen weights such that Chi-Square distance function given by

$$\sum_{h=1}^{L} \frac{\left(\Omega_h - D_h\right)^2}{D_h Q_h} \tag{4.12}$$

is minimum subject to higher level calibration equation defined as

$$\sum_{h=1}^{L} \Omega_h s_{hx}^2 = V\left(\bar{x}_{St}\right) \tag{4.13}$$

where,

$$V(\bar{x}_{St}) = \sum_{h=1}^{L} W_h^2 \frac{(1 - f_h)}{n_h} S_{hx}^2$$

is assumed to be known and $s_{hx}^2 = (n_h - 1)^{-1} \sum_{j=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ is an unbiased estimator of $S_{hx}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$. This procedure leads to a new estimator for the variance of the combined regression estimator given by

$$\hat{V}\left(\hat{Y}_{St}\right)_{\text{CLR}} = \hat{V}_{St}\left(\hat{Y}_{\text{GREG}}\right) + \hat{B}_{St}\left[V\left(\bar{x}_{St}\right) - \hat{V}\left(\bar{x}_{St}\right)\right] \tag{4.14}$$

where

$$\hat{B}_{St} = \sum_{h=1}^{L} \frac{W_h^{*2}(1 - f_h)}{n_h} Q_h s_{hx}^2 s_{\hat{e}h}^2 \Big/ \sum_{h=1}^{L} \frac{W_h^2 (1 - f_h)}{n_h} s_{hx}^4$$

denotes the combined improved estimator of regression coefficient in stratified sampling and

$$\hat{V}(\bar{x}_{St}) = \sum_{h=1}^{L} W_h^2 \frac{(1 - f_h)}{n_h} s_{hx}^2$$

is an unbiased estimator of $V(\bar{x}_{St})$. If $q_h = 1/\bar{x}_h$ and $Q_h = 1/s_{hx}^2$, then estimator (4.14) reduces to a new estimator of variance of the combined ratio estimator given by

$$\hat{V}_{St}\left(\hat{Y}_{\text{Ratio}}\right) = \sum_{h=1}^{L} \frac{W_h^2 (1 - f_h)}{n_h} s_{\hat{e}h}^2 \left(\frac{\bar{X}}{\bar{x}_{St}}\right)^2 \left\{\frac{V\left(\bar{x}_{St}\right)}{\hat{V}\left(\bar{x}_{St}\right)}\right\} \tag{4.15}$$

which is a ratio type estimator proposed by Wu (1985) for estimating variance of the combined ratio estimator but makes use of extra knowledge of the known variance of the auxiliary variable at the estimation stage. Several more new estimators can be constructed for new choices of weights q_h and Q_h.

5. A WIDER CLASS OF ESTIMATORS

If we define $u = X / \sum_{i=1}^{n} d_i x_i$ and $v = V(\hat{X}_{\text{HT}}) / \hat{V}(\hat{X}_{\text{HT}})$, then a wider class of estimators has been defined as

$$\hat{V}_{\text{SS}}\left(\hat{Y}_{\text{GREG}}\right) = \left\{\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}\left(d_i e_i - d_j e_j\right)^2\right\} H(u,v) \tag{5.1}$$

where $H(u,v)$ is a parametric function of u and v such that $H(1,1) = 1$ and which satisfies certain regularity conditions. Then all estimators obtained from the following functions,

$$H(u,v) = u^{\alpha} v^{\beta},\ H(u,v) = \frac{1 + \alpha(u - 1)}{1 + \beta(v - 1)},$$

$$H(u,v) = 1 + \alpha(u - 1) + \beta(v - 1)$$

and $H(u,v) = \{1 + \alpha(u - 1) + \beta(v - 1)\}^{-1}$ are special cases of the higher level calibration approach, where α and β are unknown parameters involved in the function $H(u,v)$. Replacing these parameters with their respective consistent estimators in the class of estimators at (5.1) leads to the same asymptotic variance as shown by Srivastava and Jhajj (1983), Singh and Singh (1984) and Mahajan and Singh (1996). The extension of present investigation to two phase sampling following Hidiroglou and Särndal (1995) is in progress.

The next section has been devoted to studying the performance of the higher order calibration approach through simulation.

6. SIMULATION STUDY

Under the simulation study, we have considered comparisons of estimators of variance of ratio estimator as well as that of regression estimator. To avoid any kind of confusion, we have redefined the estimators considered for comparison as follows:

6.1 Ratio Estimator

We have compared the estimators of the variance of the ratio estimator, given by

$$\hat{V}_1\left(\hat{Y}_{\text{RATIO}}\right) = \frac{N^2(1 - f)}{n(n - 1)} \sum_{i=1}^{n} e_i^2 \left(\frac{X}{\hat{X}}\right)^2 \tag{6.1.1}$$

with the estimator, given by

$$\hat{V}_2\left(\hat{Y}_{\text{RATIO}}\right) = \hat{V}_1\left(\hat{Y}_{\text{RATIO}}\right)\left(\frac{S_x^2}{s_x^2}\right). \tag{6.1.2}$$

6.2 Regression Estimator

We have also compared the estimators of the variance of the regression estimator, given by

$$\hat{V}_1\left(\hat{Y}_{\text{GREG}}\right) =$$

$$\frac{N^2(1 - f)}{n(n - 1)} \sum_{i=1}^{n} e_i^2 + \hat{\psi}_1\left(X - \hat{X}\right) + \hat{\psi}_2\left(X - \hat{X}\right)^2 \tag{6.2.1}$$

with the estimator, given by

$$\hat{V}_2\left(\hat{Y}_{\mathrm{GREG}}\right) = \hat{V}_1\left(\hat{Y}_{\mathrm{GREG}}\right) + \hat{\psi}_3\,(S_x^2 - s_x^2) \qquad (6.2.2)$$

where $\hat{\psi}_i, i = 1, 2, 3$ have the same meaning as defined earlier.

We have considered two types of populations *viz.* finite populations as well as infinite populations to cover almost all practical situations.

6.3 Finite Populations

In case of finite populations, we have taken a population consisting of $N = 20$ units from Horvitz and Thompson (1952). The study variable, y, is the number of households on i-th block and known auxiliary character, x, is the eye-estimated number of households on the i-th block. All possible samples of size $n = 5$ were selected by SRSWOR, which results in

$$\binom{N}{n} = 15{,}504$$

samples. From the k-th sample, the estimator

$$\hat{Y}_{\mathrm{RATIO}}|_k = \hat{Y}\left(\frac{X}{\hat{X}}\right), \text{ with } \hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

was computed. Empirical mean squared error of this estimator was computed as

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{RATIO}}\right) = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} \left[\hat{Y}_{\mathrm{RATIO}}|_k - Y\right]^2. \qquad (6.3.1)$$

For the k-th sample, the ratio type estimators of variance

$$\hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)|_k, h = 1, 2,$$

given by (6.1.1) and (6.1.2) respectively, for estimating the variance of the ratio estimator were also obtained. The bias in the h-th ratio type estimator of variance was computed as

$$B\left\{\hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)\right\} =$$

$$\binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} \hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)|_k - \mathrm{MSE}\left(\hat{Y}_{\mathrm{RATIO}}\right) \qquad (6.3.2)$$

and mean squared error was computed as

$$\mathrm{MSE}\left\{\hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)\right\} =$$

$$\binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} \left[\hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)|_k - \mathrm{MSE}\left(\hat{Y}_{\mathrm{RATIO}}\right)\right]^2. \qquad (6.3.3)$$

The percent relative efficiency of the estimator $\hat{V}_2\,(\hat{Y}_{\mathrm{RATIO}})$ with respect to $\hat{V}_1\,(\hat{Y}_{\mathrm{RATIO}})$ was calculated as

$$\mathrm{RE} =$$

$$\mathrm{MSE}\left\{\hat{V}_1\left(\hat{Y}_{\mathrm{RATIO}}\right)\right\} \times 100 / \mathrm{MSE}\left\{\hat{V}_2\left(\hat{Y}_{\mathrm{RATIO}}\right)\right\}. \qquad (6.3.4)$$

The coverage by 95% confidence intervals

$$\mathrm{CCI}\left[\hat{V}_h\left(\hat{Y}_{\mathrm{RATIO}}\right)\right]$$

for $h = 1, 2$ were calculated for h-th ratio type estimator of variance by counting the number of times the true population total, Y, falls between the limits defined as

$$\hat{Y}_{\mathrm{RATIO}}|_k \mp t_{n-h-1}\,(\alpha)\sqrt{\hat{V}_h\left(\hat{V}_{\mathrm{RATIO}}\right)|_k}. \qquad (6.3.5)$$

These results were also obtained from all possible samples of size 6 and 7 and have been presented in Table 1.

The same process was repeated for the regression estimator

$$\hat{Y}_{\mathrm{GREG}}|_k = \hat{Y} + \left(\sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2\right)\left(X - \hat{X}\right)$$

of total obtained from (1.6) under a SRSWOR design. The biases, relative efficiency and CCI were obtained by using h-th estimator of variance of the regression estimator, $\hat{V}_h\,(\hat{Y}_{\mathrm{GREG}})|_k$ for $h = 1, 2$, given by (6.2.1) and (6.2.2), respectively. The results obtained have been presented in Table 2. In addition, it was observed that for $n = 5$, 0.020% estimates of variance obtained from the estimator $\hat{V}_1\,(\hat{Y}_{\mathrm{GREG}})$ and 0.022% estimates obtained from the estimator $\hat{V}_2\,(\hat{Y}_{\mathrm{GREG}})$ were negative. Similar results were observed for more natural populations given by Cochran (1963) and Sukhatme and Sukhatme (1970). Over all, second order calibration estimators perform better than first order calibration in case of the finite populations.

In real life situations, the study variable and auxiliary variables may follow certain kinds of distributions like normal, beta or gamma *etc.* In order to see the performance of the proposed strategies under such circumstances, we generated artificial populations and considered the problem of estimation of finite population mean through simulation as follows.

Table 1
Comparison of $\hat{V}_2\left(\hat{Y}_{\text{RATIO}}\right)$ with $\hat{V}_1\left(\hat{Y}_{\text{RATIO}}\right)$ for finite populations

n	$B\left[\hat{V}_1\left(\hat{Y}_{\text{RATIO}}\right)\right]$	$B\left[\hat{V}_2\left(\hat{Y}_{\text{RATIO}}\right)\right]$	RE	$CCI\left[\hat{V}_1\left(\hat{Y}_{\text{RATIO}}\right)\right]$	$CCI\left[\hat{V}_2\left(\hat{Y}_{\text{RATIO}}\right)\right]$
5	-211.33	217.01	166.57	0.93	0.95
6	-141.92	102.00	115.06	0.91	0.92
7	-99.34	58.60	109.23	0.90	0.90

Table 2
Comparison of $\hat{V}_2\left(\hat{Y}_{\text{GREG}}\right)$ and $\hat{V}_1\left(\hat{Y}_{\text{GREG}}\right)$ for finite populations

n	$B\left[\hat{V}_1\left(\hat{Y}_{\text{GREG}}\right)\right]$	$B\left[\hat{V}_2\left(\hat{Y}_{\text{GREG}}\right)\right]$	RE	$CCI\left[\hat{V}_1\left(\hat{Y}_{\text{GREG}}\right)\right]$	$CCI\left[\hat{V}_2\left(\hat{Y}_{\text{GREG}}\right)\right]$
5	-328.49	-194.78	112.04	0.92	0.96
6	-223.92	-136.34	103.02	0.90	0.93
7	-157.88	-94.38	101.21	0.91	0.94

6.4 Infinite Populations

The size N of these populations is unknown. We generated n independent pairs of random numbers y_i^* and x_i^* (say), $i = 1, 2, ..., n$, from a subroutine VNORM with PHI = 0.7, seed(y) = 8987878 and seed(x) = 2348789 following Bratley, Fox and Schrage (1983). For fixed $S_y^2 = 50$ and $S_x^2 = 50$, we generated transformed variables,

$$y_i = 3.0 + \sqrt{S_y^2\,(1-\rho^2)}\,y_i^* + \rho\, S_y x_i^* \qquad (6.4.1)$$

and

$$x_i = 4.0 + S_x x_i^* \qquad (6.4.2)$$

for different values of the correlation coefficient ρ. For the k-th sample, the estimator

$$\hat{\bar{y}}_{\text{RATIO}}\,|_k = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right), \text{ with } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

was computed. Empirical mean squared error of this estimator was computed as

$$\text{MSE}\left(\hat{\bar{y}}_{\text{RATIO}}\right) = \frac{1}{15{,}000}\sum_{k=1}^{15{,}000}\left[\hat{\bar{y}}_{\text{RATIO}}\,|_k - \bar{Y}\right]^2. \qquad (6.4.3)$$

For the k-th sample, the ratio type estimators of variance

$$\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)|_k, h = 1, 2,$$

obtained from (6.1.1) and (6.1.2) respectively, for estimating the variance of the ratio estimator of population mean were also derived. The bias in the h-th ratio type estimator of variance was computed as

$$B\left\{\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right\} = \frac{1}{15{,}000}\sum_{k=1}^{15{,}000} \hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)|_k - \text{MSE}\left(\hat{\bar{y}}_{\text{RATIO}}\right) \qquad (6.4.4)$$

and mean squared error was computed as

$$\text{MSE}\left\{\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right\} = \frac{1}{15{,}000}\sum_{k=1}^{15{,}000}\left[\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)|_k - \text{MSE}\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right]^2. \qquad (6.4.5)$$

The percent relative efficiency of the estimator $\hat{V}_2\left(\hat{\bar{y}}_{\text{RATIO}}\right)$ with respect to $\hat{V}_1\left(\hat{\bar{y}}_{\text{RATIO}}\right)$ was calculated as

$$\text{RE} = \text{MSE}\left\{\hat{V}_1\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right\} \times 100 / \text{MSE}\left\{\hat{V}_2\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right\} \qquad (6.4.6)$$

The coverage by 95% confidence intervals

$$\text{CCI}\left[\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)\right] \text{ for } h = 1, 2$$

was calculated for h-th ratio type estimator of variance by counting the number of times the true population mean, $\bar{Y}$, falls between the limits defined as

$$\hat{\bar{y}}_{\text{RATIO}}\,|_k \mp 1.96\sqrt{\hat{V}_h\left(\hat{\bar{y}}_{\text{RATIO}}\right)|_k}. \qquad (6.4.7)$$

These results were obtained for samples of size n = 60, 80 and 100 for different values of correlation coefficient as presented in Table 3.

The same process was repeated for the regression estimator

$$\hat{\bar{y}}_{\text{GREG}}|_k = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$$

of mean obtained from (1.6) under a SRSWR design. The biases, relative efficiency and CCI were obtained by using h-th estimator of variance of the regression estimator,

$$\hat{V}_h(\hat{\bar{y}}_{\text{GREG}})|_k \text{ for } h = 1, 2,$$

derived from (6.2.1) and (6.2.2), respectively. The results obtained have been presented in Table 4. We acknowledge that it is worth while studying the proposed strategy through simulation in more detail and its application in actual practice. The empirical study was carried out in FORTRAN-77 using a PENTIUM-120.

7. CONCLUSION

Higher level calibration approach can be used if variance of the auxiliary character is known in addition to the known total of that character. The statistical package GES developed by Statistics Canada can be modified to obtain better estimators of the variance of GREG, useful for surveys where information on variance of auxiliary characters is available or can be calculated.

ACKNOWLEDGEMENTS

The authors are heartily thankful to the Associate Editor and the two learned referees for fruitful and constructive comments to bring the original manuscript in the present form. They are also grateful to Dr. M.P. Singh for his kind suggestions. The opinions and results discussed in this paper are of authors and not necessarily of their institute(s).

Table 3
Comparison of $\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})$ with $\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})$ for infinite populations

n	ρ	$B[\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})]$	$B[\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})]$	RE	$\text{CCI}[\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})]$	$\text{CCI}[\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})]$
60	0.1	13.02	10.33	188.7	0.96	0.95
	0.3	8.07	6.35	192.6	0.97	0.95
	0.5	4.33	3.37	195.9	0.96	0.96
	0.7	1.77	1.37	197.9	0.97	0.97
	0.9	0.33	0.26	197.7	0.99	0.98
80	0.1	3.27	2.91	123.2	0.94	0.93
	0.3	2.06	1.84	123.0	0.94	0.94
	0.5	1.13	1.01	122.7	0.95	0.95
	0.7	0.47	0.42	122.0	0.97	0.96
	0.9	0.08	0.08	119.1	0.98	0.97
100	0.1	0.76	0.77	106.1	0.94	0.93
	0.3	0.49	0.49	105.8	0.94	0.94
	0.5	0.27	0.27	105.3	0.95	0.95
	0.7	0.12	0.12	104.4	0.96	0.95
	0.9	0.02	0.02	102.2	0.97	0.95

Table 4
Comparison of $\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})$ with $\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})$ for infinite populations

n	ρ	$B[\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})]$	$B[\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})]$	RE	$\text{CCI}[\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})]$	$\text{CCI}[\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})]$
60	0.1	10.12	8.42	177.6	0.98	0.95
	0.3	5.06	4.33	161.5	0.97	0.95
	0.5	3.32	2.36	152.5	0.95	0.96
	0.7	0.72	0.38	151.9	0.97	0.95
	0.9	0.13	0.10	147.7	0.99	0.97
80	0.1	1.23	1.11	153.9	0.96	0.95
	0.3	1.03	1.01	143.5	0.98	0.94
	0.5	0.13	0.11	132.8	0.97	0.95
	0.7	0.07	0.06	121.6	0.97	0.95
	0.9	0.02	0.03	117.1	0.96	0.96
100	0.1	0.65	0.57	136.1	0.95	0.94
	0.3	0.39	0.32	135.1	0.94	0.94
	0.5	0.13	0.13	129.6	0.95	0.95
	0.7	0.02	0.02	114.4	0.96	0.95
	0.9	0.01	0.01	112.2	0.97	0.96

APPENDIX

This appendix has been devoted to deriving the optimum value of Ω_{ij} as given in (3.4). The Lagrange's function is given by

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\left(\Omega_{ij} - D_{ij}\right)^2}{D_{ij}Q_{ij}} -$$

$$2\lambda\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}\left(d_i x_i - d_j x_j\right)^2 - V_{\text{YG}}\left(\hat{X}_{\text{HT}}\right)\right]. \qquad (\text{A}.1)$$

On differentiating (A.1) with respect to Ω_{ij} and equating to zero, we get

$$\Omega_{ij} = D_{ij} + \lambda D_{ij} Q_{ij}\left(d_i x_i - d_j x_j\right)^2. \qquad (\text{A}.2)$$

On putting (A.2) in (3.2), we get

$$\lambda = \frac{V_{\text{YG}}\left(\hat{X}_{\text{HT}}\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\left(d_i x_i - d_j x_j\right)^2}{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij} Q_{ij}\left(d_i x_i - d_j x_j\right)^4}. \qquad (\text{A}.3)$$

On substituting (A.3) in (A.2), we get the optimum value of Ω_{ij} as given in (3.4).

REFERENCES

BRATLEY, P., FOX, B.L., and SCHRAGE, L.E. (1983). *A Guide to Simulation.* New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (second edition). New York: John Wiley and Sons.

DAS, A.K., and TRIPATHI, T.P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhyā*, 40(C), 139-148.

DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, 32, 209 - 226.

GARCIA, M.R., and CEBRIAN, A.A. (1996). Repeated substitution method: The ratio estimator for the population variance. *Metrika*, 43, 101-105.

HIDIROGLOU, M. A., and SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* Volume II, 873-878.

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

ISAKI, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78(381), 117-123.

MAHAJAN, P.K., and SINGH, S. (1996). On estimation of total in two stage sampling. *Journal of Statistical Research*, 30, 127-131.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.

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

SAXENA, S.K., NIGAM, A.K., and SHUKLA, N.D. (1995). Variance estimation for combined ratio estimator. *Sankhyā*, 57(B), 85-92.

SHAH, D.N., and PATEL, P.A. (1996). Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling. *The Canadian Journal of Statistics*, 24(3), 373-384.

SINGH, P., and SRIVASTAVA, S.K. (1980). Sampling scheme providing unbiased regression estimators. *Biometrika*, 67, 205-209.

SINGH, R.K., and SINGH, G. (1984) A class of estimators with estimated optimum values in sample surveys. *Statistics & Probability Letters*, 2, 319-321.

SINGH, S., and SINGH, S. (1988). Improved estimators of K and B in finite populations. *Journal of the Indian Society of Agricultural Statistics*, 121-126.

SINGH, S., MANGAT, N.S., and MAHAJAN, P.K. (1995). General class of estimators. *Journal of the Indian Society of Agricultural Statistics*, 47(2), 129-133.

SRIVASTAVA, S.K. (1971). A generalized estimator for the mean of finite population using multi-auxiliary information. *Journal of the American Statistical Association*, 66, 404-407.

SRIVASTAVA, S.K., and JHAJJ, S.K (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā* 42(C), 87-96.

SRIVASTAVA, S.K., and JHAJJ, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, 68, 341-343.

SRIVASTAVA, S.K., and JHAJJ, H.S. (1983). A class of estimators of estimators of the population mean using multi-auxiliary information. *Calcutta Statistical Association Bulletin 32*, 47-56.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications.* Iowa: Iowa State University Press.

SWAIN, A.K.P.C., and MISHRA, G. (1992). Unbiased estimators of finite population variance using auxiliary information. *Metron*, 201-215.

WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.

WU, C.F.J. (1985). Variance estimation for combined ratio and combined regression estimators. *Journal of the Royal Statistical Society*, 47(B), 147-154.

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

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 51-55
Statistics Canada

Logistic Generalized Regression Estimators

RISTO LEHTONEN and ARI VEIJANEN[1]

ABSTRACT

In this paper we study the model-assisted estimation of class frequencies of a discrete response variable by a new survey estimation method, which is closely related to generalized regression estimation. In generalized regression estimation the available auxiliary data are incorporated in the estimation procedure by a linear model fit. Instead of using a linear model for the class indicators, we describe the joint distribution of the class indicators by a multinomial logistic model. Logistic generalized regression estimators are introduced for class frequencies in a population and domains. Monte Carlo experiments were carried out for simulated data and for real data taken from the Labour Force Survey conducted monthly by Statistics Finland. The logistic generalized regression estimation yielded better results than the ordinary regression estimation for small domains and particularly for small class frequencies.

KEY WORDS: Auxiliary information; Class frequencies; Generalized linear models; Labour force survey; Model-assisted estimation; Regression estimators.

1. INTRODUCTION

Consider the estimation of class frequencies of a discrete response variable in a sample survey. The number of individuals in a class equals the class indicator's sum over the population, the total of the indicator. Therefore, the problem can be solved by methods designed for the estimation of population totals. To improve the accuracy of the estimation, a survey statistician often makes use of the available auxiliary data. If the expectation of the response variable can be assumed to depend linearly on the auxiliary variables as can be the case for continuous response variables, it is advisable to use the generalized regression estimator (Särndal, Swensson and Wretman 1992; Estevao, Hidiroglou and Särndal 1995). Generalized regression estimation can improve the efficiency and reduce the bias due to unit nonresponse if the auxiliary variables correlate strongly with the response variable.

From a modeler's perspective, a linear model is quite restrictive and might not be the best choice for binary response variables, such as employment status of a person (employed, unemployed), or more generally for discrete response variables, such as a person's status in the labour market (employed, unemployed, not in labour force). For such variables we introduce a class of logistic generalized regression estimators based on a multinomial logistic model describing the joint distribution of the class indicators. The motivation for the selection of this specific model type thus is similar to that used in the context of generalized linear models (McCullagh and Nelder 1989).

The parameters of the logistic model are here estimated by maximizing a sample-based weighted loglikelihood, the Horvitz-Thompson estimator of the population loglikelihood function (Godambe and Thompson 1986; Nordberg 1989; Skinner, Holt and Smith 1989; Särndal *et al.* 1992, p. 517).

As an application, we consider the estimation of the unemployment rate in the Labour Force Survey conducted monthly by Statistics Finland. Administrative records indicating whether a person is registered jobseeker in local employment office are available as register-based auxiliary data, and these records were merged with the survey data on individual basis using personal identification numbers which are unique in both data sources. The corresponding auxiliary variable correlates strongly with the survey measurement on person's unemployment. Thus, improvement in efficiency and reduction of bias can be expected by making use of these administrative data in the estimation procedure. Additional auxiliary data (sex, age, regional data) were gathered from the Population Register. Also these auxiliary data were merged with the survey data on individual basis.

The properties of the generalized regression estimators were studied by Monte Carlo simulation methods where SRSWOR samples were repeatedly drawn from a population constructed from the Labour Force Survey data. We use incomplete poststratification or raking based on a main effects ANOVA model. The experiments indicate that the logistic formulation yields better results than the linear formulation for small domains. We obtained good results also when there was only one continuous auxiliary variable.

This paper is organized as follows. Section 2 defines the multinomial logistic model and basic concepts used. In Section 3 we introduce generalized regression estimators of class frequencies in a population and domains, and discuss the estimation of the model parameters by weighted loglikelihood. Variance estimators are presented. Monte Carlo experiments are discussed in Section 4. Conclusions are drawn in Section 5.

[1] Risto Lehtonen and Ari Veijanen, Statistics Finland, P.O. Box 5A, FIN-00022 Statistics Finland, Finland.

2. MODEL

Consider discrete m-valued random variables Y_k associated with N elements k in a finite population U. We observe their realized values y_k only in a sample $s \subset U$ of size n. Our goal is to estimate the frequency distribution of the y_k's in the population; in classification problems, we estimate the class proportions. Suppose we know the vector of auxiliary variables $\boldsymbol{x}_k$ for every element in the population. We impose a multinomial logistic model

$$P\{Y_k = i\} = \frac{\exp\{\boldsymbol{x}_k'\boldsymbol{\beta}_i\}}{\sum_{r=1}^{m} \exp\{\boldsymbol{x}_k'\boldsymbol{\beta}_r\}} \qquad (i = 1, 2, \ldots, m) \qquad (1)$$

and assume that the Y_k's are conditionally independent given the $\boldsymbol{x}_k$'s. In the binary case, this is the model used in logistic regression. The parameter vector $\boldsymbol{\beta}$ is composed of vectors $\boldsymbol{\beta}_i\,(i = 1, 2, \ldots, m)$ with components $\beta_{ij}\,(j = 1, 2, \ldots, q)$. The parameters are assumed identifiable, that is, no two parameter values yield identical probabilities (1) for every k. This implies that the auxiliary variables $x_{kj}\,(j = 1, 2, \ldots, q)$ are linearly independent. To avoid identifiability problems, we set $\boldsymbol{\beta}_1 = 0$. It is straightforward to generalize (1) so that different auxiliary variables can be assigned for the m classes (Lehtonen and Veijanen 1998).

The sampling design specifies the inclusion probabilities of population elements. The k-th element is drawn with inclusion probability π_k and elements k and p are simultaneously in the sample s with probability $\pi_{kp} > 0$ $(\pi_{kk} = \pi_k)$. As usual, the sample membership indicators $I_k = I\{k \in s\}$ are assumed conditionally independent of the Y_k's given the $\boldsymbol{x}_k$'s, but the inclusion probabilities may correlate with the auxiliary variables.

Under unit nonresponse, if element k responds with probability θ_k independently of the I_p's and Y_p's $(p \in U)$, then we substitute $\pi_k\theta_k$ for π_k. Correspondingly, π_{kp} is replaced by $\pi_{kp}\,\theta_k\,\theta_p$ when the elements respond independently of each other.

3. LOGISTIC GENERALIZED REGRESSION ESTIMATION

3.1 Definition of LGREG

To estimate the frequency distribution of the y_k's, we define class indicators $Z_{ki} = I\{Y_k = i\}$ with realizations z_{ki} and estimate the totals $t_i = \sum_{k \in U} z_{ki}$. The Horvitz-Thompson (HT) estimator of t_i is $\hat{t}_i^{\mathrm{HT}} = \sum_{k \in s} a_k z_{ki}$, where the sampling weights are $a_k = 1/\pi_k$. Generalized regression estimation (GREG) is assisted by a regression model $Z_{ki} = \boldsymbol{x}_k'\,\boldsymbol{\beta}_i^G + \varepsilon_{ki}$ with $\mathrm{Var}(\varepsilon_{ki}) = \sigma_{ki}^2$ (Särndal *et al.* 1992; Estevao *et al.* 1995). The parameter $\boldsymbol{\beta}_i^G$ is estimated by

$$\hat{\boldsymbol{\beta}}_i^G = \left(\sum_{k \in s} a_k \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_{ki}^2} \right)^{-1} \left(\sum_{k \in s} a_k \frac{\boldsymbol{x}_k z_{ki}}{\sigma_{ki}^2} \right) \quad (i = 1, 2, \ldots, m) \quad (2)$$

and the fitted values $\hat{z}_{ki} = \boldsymbol{x}_k'\,\hat{\boldsymbol{\beta}}_i^G$ are incorporated in the GREG estimator

$$\hat{t}_i^G = \sum_{k \in U} \hat{z}_{ki} + \sum_{k \in s} a_k (z_{ki} - \hat{z}_{ki}) \quad (i = 1, 2, \ldots, m). \qquad (3)$$

The selection of a linear model for a GREG estimator (3) is fully justified for a continuous response variable. For binary measurements Z_{ki}, a linear model might be unrealistic. Ordinarily, we would prefer a logistic model to a linear one. In the logistic formulation, the predicted value always lies in [0,1], whereas in the linear formulation, the predicted value can exceed these natural limits. If the probability of $Z_{ki} = 1$ is close to 0 or 1, then the two models yield different results. Moreover, when there are $m > 2$ classes, it appears sensible to describe the joint distribution of the Z_{ki}'s $(i = 1, 2, \ldots, m)$ by the multinomial logistic model (1). To apply the model (1) in generalized regression estimation, we estimate the expectations $\mu_{ki} = E(Z_{ki} \mid \boldsymbol{x}_k; \boldsymbol{\beta}) = P\{Y_k = i \mid \boldsymbol{x}_k; \boldsymbol{\beta}\}$ by

$$\hat{\mu}_{ki} = P\{Y_k = i \mid \boldsymbol{x}_k; \hat{\boldsymbol{\beta}}\} = \frac{\exp\{\boldsymbol{x}_k'\,\hat{\boldsymbol{\beta}}_i\}}{1 + \sum_{r=2}^{m} \exp\{\boldsymbol{x}_k'\,\hat{\boldsymbol{\beta}}_r\}},$$

which depend nonlinearly on the auxiliary variables. We define a logistic generalized regression (LGREG) estimator by

$$\hat{t}_i = \sum_{k \in U} \hat{\mu}_{ki} + \sum_{k \in s} a_k (z_{ki} - \hat{\mu}_{ki}) \quad (i = 1, 2, \ldots, m). \qquad (4)$$

The GREG and LGREG estimators (3) and (4) include a sum of predicted values over the population. However, it is not actually necessary to have information about the $\boldsymbol{x}_k$'s for every element in the population U. In GREG (3), it is enough to know the auxiliary totals $\sum_{k \in U} \boldsymbol{x}_k$, because (3) can also be expressed in the form $\hat{t}_i^G = \hat{t}_i^{\mathrm{HT}} + (\sum_{k \in U} \boldsymbol{x}_k - \sum_{k \in s} a_k \boldsymbol{x}_k)'\hat{\boldsymbol{\beta}}_i^G$. For the special case of complete poststratification, the information required in LGREG is similar to that needed in GREG. For other cases, such as incomplete poststratification, we cannot compute $\sum_{k \in U} \hat{\mu}_{ki}$ in (4) without knowing the frequency of each value of $\boldsymbol{x}_k$ in the population. For example, if we have two discrete auxiliary variables, then in GREG we need the marginal frequencies, but in LGREG we need the cell frequencies.

In addition to estimates for the whole population, estimates are usually calculated for subpopulations. The population U is partitioned into *domains* $U_{(d)} \subset U$ of size

$N_{(d)}$. The set s of respondents is composed of corresponding subsets $s_{(d)} = s \cap U_{(d)}$ with $n_{(d)}$ elements. As in GREG estimation (Särndal *et al.* 1992), we apply LGREG estimator

$$\hat{t}_{(d)i} = \sum_{k\in U_{(d)}} \hat{\mu}_{ki} + \sum_{k\in s_{(d)}} a_k (z_{ki} - \hat{\mu}_{ki}). \qquad (5)$$

These estimators are additive: $\sum_i \hat{t}_{(d)i} = N_{(d)}$. If we combine two nonoverlapping domains d_1 and d_2, the LGREG estimate for $d = d_1 \cup d_2$ is $\hat{t}_{(d)i} = \hat{t}_{(d_1)i} + \hat{t}_{(d_2)i}$. Hence, $\sum_d \hat{t}_{(d)i} = \hat{t}_i$ for nonoverlapping domains and $\sum_i \hat{t}_i = N$.

In generalized regression estimation, an estimate (3) or (4) can be negative, when negative residuals coincide with large values of a_k. Negative GREG estimates become more common, as the number of auxiliary variables increases (Chambers 1996). In LGREG estimation, in contrast, this is not so, because $\hat{\mu}_{ki}$ is bounded by the model formulation. In our experiments, LGREG estimates were negative only for small domains in certain cases. In many cases, LGREG estimate equals the sum of estimated expectations and then it is always positive (see Section 3.2).

If the model (1) includes an auxiliary indicator variable, its total over the population is exactly estimated by LGREG. This calibration property is desirable in many applications.

3.2 ML Estimation by π-Weighted Loglikelihood

We estimate the parameter $\boldsymbol{\beta}$ in the model (1) by maximizing a π-weighted loglikelihood

$$L_s(\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_m) =$$

$$\sum_{k\in s} \pi_k^{-1} \left\{ I\{Y_k = 1\} \log\left(1 - \sum_{i=2}^{m} \mu_{ki}\right) + \sum_{i=2}^{m} I\{Y_k = i\} \log \mu_{ki} \right\}$$

(Godambe and Thompson 1986; Nordberg 1989; Särndal *et al.* 1992, p. 517). In general, we maximize the likelihood function numerically by appropriate numerical methods such as a Newton-Raphson algorithm.

It can be shown that for complete poststratification, the fitted values $\hat{z}_{ki}$ in GREG are equal to the estimates $\hat{\mu}_{ki}$ in LGREG. Thus, when there are no missing cells in complete poststratification, the GREG and LGREG estimators are identical (Lehtonen and Veijanen 1998). This does not hold for other models such as incomplete poststratification.

The LGREG estimator (4) has two parts: a sum of estimated expectations over the population and an adjustment term $\sum_{k\in s} a_k(z_{ki} - \hat{\mu}_{ki})$. It can be shown that if the model contains an intercept, the adjustment term vanishes and the frequency t_i is estimated by $\sum_{k\in U} \hat{\mu}_{ki}$ (Lehtonen and Veijanen 1998).

In our experiments, we apply a ratio estimator $\hat{R} = \hat{t}_i / (\hat{t}_i + \hat{t}_j)$. Its variance is estimated by Taylor linearization techniques (Särndal *et al.* 1992, p. 179):

$$\hat{V}(\hat{R}) = \frac{1}{(\hat{t}_i + \hat{t}_j)^2} \left[(1 - \hat{R})^2 \hat{C}_{ii} + 2\hat{R}(\hat{R} - 1)\hat{C}_{ij} + \hat{R}^2 \hat{C}_{jj} \right] \qquad (6)$$

where C_{ij}, the covariance of $\hat{t}_i$ and $\hat{t}_j$, is estimated by

$$\hat{C}_{ij} = \sum_{k,p\in s} \frac{\Delta_{kp}}{\pi_{kp}} \frac{e_{ki}\, e_{pj}}{\pi_k\, \pi_p}. \qquad (7)$$

In (7), $e_{ki} = z_{ki} - \hat{\mu}_{ki}$ and $\Delta_{kp} = \mathrm{Cov}(I_k, I_p) = \pi_{kp} - \pi_k \pi_p$. Similar derivations hold for the corresponding domain estimators.

4. EXPERIMENTS

4.1 Details of Simulation Studies

In all the simulation experiments, $K = 1{,}000$ samples were drawn from a population with simple random sampling without replacement (SRSWOR). Monte Carlo means and standard errors of the estimates were calculated from the simulated samples. The design effect for an estimator $\hat{t}_{(d)i}$ was calculated as a ratio of estimated variances: $\mathrm{Deff}(\hat{t}_{(d)i}) = \hat{V}_{mc}(\hat{t}_{(d)i}) / \hat{V}_{mc}(\hat{t}^{\mathrm{HT}}_{(d)i})$, where $\hat{V}_{mc}(\hat{t}^{\mathrm{HT}}_{(d)i})$ denotes the Monte Carlo variance estimate of the HT estimator (Lehtonen and Pahkinen 1996). We measured the overall accuracy of domain estimates by the mean absolute relative domain error over D domains and K samples s_j:

$$\mathrm{MARDE}(i) = \frac{1}{D} \sum_{p=1}^{D} \frac{1}{K} \sum_{j=1}^{K} \frac{100 \left| \hat{t}_{(d_p)i}(s_j) - t_{(d_p)i} \right|}{t_{(d_p)i}}.$$

In the GREG estimates (2), the variance was a constant $\sigma^2_{ki} = \sigma^2$, which cancelled out. For LGREG, domain frequencies were estimated by (5) and variances by (7). For GREG and HT, see Särndal *et al.* (1992, p. 401). Confidence intervals for the frequencies were computed as if the class indicators were independent. At the nominal significance level of 95%, an acceptable coverage rate lies in [93.65%, 96.35%] for $K = 1{,}000$ simulated samples.

4.2 An Experiment With Simulated Data

To compare LGREG with GREG, we simulated a data set, in which the auxiliary variable X was a continuous random variable uniformly distributed in (-3,3). The variable of interest, Y, representing three classes followed distribution (1) specified by $\mathbf{x}_k'\boldsymbol{\beta}_1 = 0$, $\mathbf{x}_k'\boldsymbol{\beta}_2 = 3X_k - 1$, and $\mathbf{x}_k'\boldsymbol{\beta}_3 = -2X_k$ for $N = 10{,}000$ elements ($k = 1, 2, ..., N$). A thousand samples of size n = 1,000 were independently drawn with SRSWOR. X_k and X_k^2 were used as auxiliary variables. All the estimators appeared unbiased (Table 1). The variance estimates had empirical bias smaller than 3% and standard deviation smaller than 5%.

Table 1

The design effects (Deff) for class frequency estimators and the empirical coverage rates (CR) (%) of 95% confidence intervals for classes i = 1, 2, 3

Method	Deff			CR		
	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$
HT	1	1	1	95.2	95.3	94.7
GREG	0.93	0.55	0.57	95.0	94.3	95.6
LGREG	0.89	0.45	0.50	94.9	93.7	95.3

The best results were obtained by LGREG, probably due to the fact that the proportional frequencies of classes varied greatly over the range of the auxiliary variable. The probability of each class was such a function of the continuous auxiliary variable that a linear regression model did not fit the data well.

4.3 An Experiment With the Finnish Labour Force Survey Data

4.3.1 Constructed Population

We studied the estimation of the unemployment rate using the Finnish Labour Force Survey (LFS) data of three consecutive months of the year 1994. The constructed population consisted of 33,329 individuals. From the Population Register we obtained, for each population member, age class (15-24, 25-34, 35-44, 45-54, and 55-64 years), sex and region (three areas). A jobseeker indicator was obtained from the register maintained by Ministry of Labour showing which individuals were registered as unemployed jobseekers. The time lag in this administrative data source is about two weeks. It can thus be expected that the proportion of persons with changes in the actual labour market status is small within this short time interval. It should be noticed that the register-based jobseeker status is defined differently from the employment status measured in the Labour Force Survey. The survey measurement is based on a standard International Labour Office (ILO) definition. All these auxiliary data were merged with the survey data on individual basis.

The nonresponse rate varied by jobseeker status so that among registered jobseekers the rate was 11.4% whereas for the others the rate was 7.6%. The probability of nonresponse was modeled by a logistic ANOVA model and the ML estimates of nonresponse rates (ranging from 2.9% to 22.8%) were used as a nonresponse model in simulations.

For simulation experiments, we constructed an artificial population consisting of N = 30,835 persons. Employment status was defined by three classes: "employed", "unemployed", and "not in labour force" with population frequencies t_1 = 17,373, t_2 = 4,433, and t_3 = 9,029, respectively. The unemployment rate was defined by $R = t_2/(t_1 + t_2) = 20.33\%$. As domains we used the cells in the crosstabulation of age classes, sex, and the register-based unemployment status.

From the artificial population, K = 1,000 independent random samples of size n = 1,000 persons were drawn with simple random sampling without replacement. In each sample, nonresponse was simulated by the nonresponse model fitted to the original population. The response probabilities were then estimated from each sample by logistic regression with the same ANOVA model as in the nonresponse model. We multiplied each probability π_k by the estimated response probability.

Three models were used to compare LGREG with GREG. The components of x_k were dummies corresponding to age (5 classes), sex, region (3 areas) and jobseeker status. In incomplete poststratification, or raking, a main effects ANOVA model was based on classified auxiliary variables. We compared models with and without the jobseeker indicator. The third model also included a fourth-order polynomial of age.

4.3.2 Results

Incorporating no auxiliary information, HT estimators had usually larger variance than the generalized regression estimators (Table 2). Both generalized regression estimators based on a raking model with age, sex, and region yielded some improvement over the HT estimates. Much better results were obtained by models including the jobseeker indicator, which correlates more strongly (r = 0.83) with the ILO unemployment indicator than the other auxiliary variables. Thus these auxiliary data improve the efficiency of estimation (*cf.* Djerf 1997).

Table 2

Properties of unemployment rate estimates ($\hat{R}(\%)$) for the raking model (R) and the model including age polynomial (P), with (E) or without (N) the jobseeker indicator. SD denotes the standard deviation and CR (%) denotes the coverage rate of 95% confidence intervals

Model	Method	$\hat{R}$	Bias	SD	Deff	CR	MARDE
	HT	20.32	-0.0081	1.461	1	95.7	35.28
RN	GREG	20.30	-0.0262	1.454	0.995	95.3	46.03
RN	LGREG	20.31	-0.0229	1.454	0.995	95.3	45.93
RE	GREG	20.30	-0.0244	0.895	0.612	96.0	35.74
RE	LGREG	20.29	-0.0419	0.901	0.617	94.8	34.80
PE	GREG	20.30	-0.0259	0.887	0.607	95.6	35.41
PE	LGREG	20.29	-0.0421	0.896	0.613	95.1	34.76

Table 3
Mean absolute relative domain errors (MARDE) and mean coverage rates (CR) (%) of 95% confidence intervals for estimated class frequencies in domains with true frequency $t_{(d)i}$ (i = 1, 2, 3) (a) smaller than 100, and (b) at least 100. The model included the age polynomial

	Method	MARDE			CR		
		$\hat{t}_{(d)1}$	$\hat{t}_{(d)2}$	$\hat{t}_{(d)3}$	$\hat{t}_{(d)1}$	$\hat{t}_{(d)2}$	$\hat{t}_{(d)3}$
(a)	GREG	96.92	67.36	121.95	88.2	77.8	84.6
	LGREG	80.28	67.20	104.05	83.9	76.5	51.7
(b)	GREG	6.95	12.31	14.35	94.1	85.9	93.7
	LGREG	6.88	12.34	14.29	93.9	85.4	93.3

The differences between GREG and LGREG were small at the population level (Table 2). LGREG was never inferior to GREG. Domain totals, especially in small domains, were more accurately estimated by LGREG than by GREG (Table 3). When the model included the age as a continuous auxiliary variable, the standard deviation of the unemployment rate estimate was smaller for LGREG than for GREG in 19 of 20 domains. Unfortunately, the confidence intervals obtained by LGREG were often too narrow due to small variance estimates (Table 3).

5. SUMMARY

We introduce a new approach to the model-assisted estimation of population class frequencies of a discrete response variable in survey sampling. Our logistic generalized regression estimation (LGREG) is based on a multinomial logistic model, which might be more realistic for class indicators than the linear model normally used in generalized regression estimation (GREG). LGREG and GREG estimators yield identical results for complete poststratification, but differ for other models such as raking. As compared with GREG, LGREG usually requires more auxiliary information, not only the auxiliary totals. Nevertheless, LGREG appears preferable to GREG when the class probabilities vary greatly over the range of continuous auxiliary variables and when we need estimates for small domains, particularly in the presence of small class frequencies.

ACKNOWLEDGEMENTS

We are grateful to Prof. Carl-Erik Särndal, University of Montreal, for his comments on the previous version of the manuscript. Detailed comments given by an Associate Editor and two referees were very helpful. Thanks are also due to Mr. Timo Koskimäki, Statistics Finland, for providing us with the Labour Force Survey data and to Mr. Kari Djerf for helpful comments.

REFERENCES

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

DJERF, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, 13, 29-39.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.

LEHTONEN, R., and PAHKINEN, E.J. (1996). *Practical Methods for Design and Analysis of Complex Surveys*. Revised Edition. Chichester: John Wiley & Sons.

LEHTONEN, R., and VEIJANEN, A. (1998). On Multinomial Logistic Generalized Regression Estimators. Jyreäskylä. Preprints from the Department of Statistics, University of Jyreäskylä, 22.

McCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*. Second Edition. London: Chapman and Hall.

NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

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

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 57-67
Statistics Canada

Confidence Intervals for Domain Parameters When the Domain Sample Size is Random

ROBERT J. CASADY, ALAN H. DORFMAN and SUOJIN WANG[1]

ABSTRACT

Let A be a population domain of interest and assume that the elements of A cannot be identified on the sampling frame and the number of elements in A is not known. Further assume that a sample of fixed size (say n) is selected from the entire frame and the resulting domain sample size (say n_A) is random. The problem addressed is the construction of a confidence interval for a domain parameter such as the domain aggregate $T_A = \sum_{i\in A} x_i$. The usual approach to this problem is to redefine x_i, by setting $x_i = 0$ if $i \notin A$. Thus, the construction of a confidence interval for the domain total is recast as the construction of a confidence interval for a population total which can be addressed (at least asymptotically in n) by normal theory. As an alternative, we condition on n_A and construct confidence intervals which have approximately nominal coverage under certain assumptions regarding the domain population. We evaluate the new approach empirically using artificial populations and data from the Bureau of Labor Statistics (BLS) Occupational Compensation Survey.

KEY WORDS: Bayes method; Conditioning; Establishment surveys; Simple random sampling; Stratification; Survey methods.

1. INTRODUCTION

In sampling from a finite population, we often are interested in the estimation of totals, means, or other quantities, for parts of that population, usually referred to as domains. Such domains are not explicitly listed in the frame, the number of items that will occur in the survey is not known in advance, and often enough, we do not even know the number of their elements in the population. For example, we might sample schoolchildren for certain medical problems, and then wish to know the mean blood pressure of those children who are underweight. The class of underweight children would constitute a domain. The only information we have as to whether or not a child is underweight is likely to be among the sampled children; if so, then this would be a case where the domain is not explicitly listed on the frame.

An essential part of the inference process is the estimation of the precision of our estimators; this is typically given by an estimated standard deviation, coefficient of variation, or confidence interval. The notion of a valid confidence interval underlies whatever measure of precision we use. All confidence intervals have, by construction, a stated "nominal" confidence level. A valid confidence interval is a confidence interval with actual coverage matching the nominal coverage. The actual coverage may be determined theoretically or by empirical work mimicking the practical circumstances in which the confidence interval would be used. If a standard deviation is not such as to give rise to a valid confidence interval, then the standard deviation needs to be regarded as misleading.

In the case of estimates for domains, confidence intervals constructed along traditional lines can lead to serious under-coverage, a fact not always appreciated in the literature. We refer to this as the domain problem. The present paper addresses this problem by a somewhat complex methodology involving Bayesian ideas, which, however, leads to a rather simple practical solution, improving on current methodology. The main change in method lies in replacing the standard normal statistic used in the construction of confidence intervals, with a Student's t-statistic having degrees of freedom that depend on the number and configuration of the domain items in the sample.

We shall focus on domain totals and domain means for the two common cases of simple random sampling and stratified random sampling. In the case of simple random sampling, it turns out that standard methods are satisfactory for the mean; however, for the total, coverage can be lower than nominal but not usually worrisome. For stratified random sampling, confidence intervals for both the mean and the total pose serious difficulties with regard to coverage level. In this case, the new methodology is augmented by use of a well known approximation due to Satterthwaite (1946). Alternate approaches to ours, also using this approximation, may be found in Johnson and Rust (1993) and Kott (1994).

An outline of the paper is as follows: In Section 2, to introduce ideas, we consider the case of the total in simple random sampling, using it to illustrate the standard approach for domain estimation, the coverage problem to which this gives rise, and the approach here taken to rectify the difficulty. Section 3 describes the extension to stratified random sampling. Section 4 states our conclusions.

[1] Robert J. Casady and Alan H. Dorfman, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001, U.S.A.; Suojin Wang, Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

2. THE CASE OF SIMPLE RANDOM SAMPLING

2.1 Standard Method

The standard approach to domain estimation is well described in Särndal, Swensson, and Wretman (1992); Sections 3.3, 5.8, and Chapter 10) (henceforth SSW). Their approach is general. Here we paraphrase it for the case of simple random sampling, and, by mild extension, for stratified random sampling as well, and focus on the domain total.

Let x_i be the value of the characteristic of interest for the i-th ($i = 1, 2, ..., N$) element of the population and let A be a domain of interest. We shall consider only the case where the elements of A cannot be identified on the frame and the number N_A of elements in A is not known; the case where N_A is known is fully treated in SSW. It is assumed that any element of A included in a sample can be identified. The problem is to construct a confidence interval for the domain total, $T_A = \sum_{i \in A} x_i$, based on a sample of n elements selected from the entire frame.

Explicitly (as in SSW, Section 3.3) or implicitly (as in SSW, Section 10.3) the standard approach to this problem is to redefine x_i, by setting $x_i = 0$ if $i \notin A$, which forces the population total $T = \sum_{i=1}^{N} x_i$ to be equal to T_A. Thus, the construction of a confidence interval for the domain total is recast as the construction of a confidence interval for a population total. In what follows it is assumed that the x_i's have been redefined as above. We shall also assume, here and throughout this paper, that n is sufficiently large and n/N sufficiently small that second order terms can be ignored. Define the additional population parameters,

$\bar{X} = T/N$ = population mean,

$S^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2/N$ = population variance, and

$p_A = N_A/N$ = proportion of population in A.

Then

(1) $\hat{T}_A = (N/n) \sum_{i=1}^{n} x_i$, $\bar{x} = \sum_{i=1}^{n} x_i/n = \hat{T}_A/N$, $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$, and $\hat{p}_A = n_A/n$ (where n_A is the number of sample elements in A) are unbiased for the corresponding population parameters,

(2) $E(\hat{T}_A) = T_A$,

(3) $\text{var}(\hat{T}_A) = N^2 S^2/n$,

(4) $\sqrt{n}(\hat{T}_A - T_A)/(NS) \xrightarrow{d} N(0, 1)$, and

(5) s^2 is consistent for S^2.

It follows that $\sqrt{n}(\hat{T}_A - T_A)/(Ns) \xrightarrow{d} N(0, 1)$, so, when n is "sufficiently large", appropriate values from the normal distribution can be used to construct confidence intervals for T_A, as noted by SSW, p. 391.

The proportion of the population in A^c is $1 - p_A$ and $x_i = 0$ for $i \in A^c$; therefore, when p_A is small and the values of the x_i's for $i \in A$ are concentrated away from zero, the convergence in distribution in (4) can be slow. Consequently, the distribution of $\sqrt{n}(\hat{T}_A - T_A)/Ns$ can deviate from normal even for what are usually considered to be moderate to large values of n. The simulation study in Section 2.5 illustrates this.

For the case of stratified random sampling, confidence interval coverage for domain quantities using standard methods can be poor. Dorfman and Valliant (1993) noted the problem in their study of wage distributions for domains consisting of workers in specific occupational groups. Preliminary empirical work by the authors indicated that supposed 95% confidence intervals for total workers and total wages for occupation based domains typically provided only 75% to 85% coverage even for a large total sample size ($n = 353$ establishments). These results are verified as part of the empirical work described in Section 3. Furthermore, their work indicated that the distribution of $\hat{T}_A - T_A$ was strongly dependent on the realized value of n_A, which suggested that some type of "conditional" confidence interval should be considered. It seems desirable to establish methodology for the construction of conditional (on n_A or equivalently $\hat{p}_A$) confidence intervals for T_A, which provide nominal, or near nominal, coverage regardless of the realized value of the domain sample size. Inference conditional on sample size is discussed in SSW, Section 10.4, but only for the case of known N_A; we are concerned throughout this paper with the case of unknown N_A.

2.2 Definitions and Notation

We define the following parameters and estimators:

Domain parameters:

$\mu_A = T_A/N_A$ = domain mean,

$\sigma_A^2 = \sum_{i \in A} (x_i - \mu_A)^2/N_A$ = variance of population elements in A.

Domain estimators:

$\hat{N}_A = \hat{p}_A N$,

$\hat{\mu}_A = \sum_{i=1}^{n_A} x_i/n_A = \hat{T}_A/\hat{N}_A$ (only defined for $n_A \geq 1$), and

$\hat{\sigma}_A^2 = \sum_{i=1}^{n_A} (x_i - \hat{\mu}_A)^2/(n_A - 1)$ (only defined for $n_A \geq 2$).

In what follows it is understood that $n_A \geq 2$ (or equivalently $\hat{p}_A \geq 2/n$) unless specifically stated otherwise. At $n_A = 1$ or 0, it is preferable to supply an "insufficient information" tag, rather than attempt inference. The relationships given below follow directly from the definitions:

$T_A = N p_A \mu_A$ and $\hat{T}_A = N \hat{p}_A \hat{\mu}_A$,

$\bar{X} = p_A \mu_A$ and $\bar{x} = \hat{p}_A \hat{\mu}_A$,

$S^2 = p_A(1 - p_A)\mu_A^2 + p_A \sigma_A^2$

and

$$s^2 = \frac{n}{n-1}\hat{p}_A\left(1 - \hat{p}_A\right)\hat{\mu}_A^2 + \frac{n\hat{p}_A - 1}{n-1}\hat{\sigma}_A^2. \qquad (1)$$

Also, it is straightforward to verify that

$$\left(\sqrt{n}/N\right)\left(\hat{T}_A - T_A\right) = \sqrt{n}\mu_A\left(\hat{p}_A - p_A\right) + \sqrt{\hat{p}_A}\sigma_A Z, \qquad (2)$$

where $Z = \sqrt{n\hat{p}_A}(\hat{\mu}_A - \mu_A)/\sigma_A$. Thus, conditionally on $\hat{p}_A$, $\hat{T}_A$ is biased for T_A, and if, for example, we assume an underlying normality, and standardize $(\sqrt{n}/N)(\hat{T}_A - T_A)$ by the corresponding conditional variance, we will get a non-central t-distribution with unknown non-centrality parameter proportional to $\sqrt{n}\mu_A(\hat{p}_A - p_A)$, providing little basis for (conditional) sound inference. This is the problem which the discussions in the next sections attempt to address.

We remark that in estimating the mean μ_A by $\hat{\mu}_A$, the bias is zero, and the problem of the preceding paragraph does not arise. This is the reason that, in simple random sampling, standard inference for means is sound, at least when the domain variates are normally distributed.

2.3 General Methodology for Confidence Intervals

Let $\hat{\theta} = (\hat{T}_A - T_A)/s_{\hat{T}_A}$, where $s^2_{\hat{T}_A}$ is an estimator (to be specified) of the (conditional or unconditional) variance of the total. Assume that the form of the conditional (on $\hat{p}_A$) distribution function of $\hat{\theta}$, say $H(\cdot\,|\,\hat{p}_A; p_A, \mu_A, \sigma^2_A)$, is known where p_A, μ_A and σ^2_A represent unknown parameters. In order to construct a conditional equal tailed $(1 - \alpha) \times 100\%$ confidence interval (CI) for T_A, we define an upper critical value

$$c_u \equiv c_u(\alpha, \hat{p}_A; p_A) = -\inf\left\{x \,|\, H\left(x\,|\,\hat{p}_A; p_A\right) \geq \alpha/2\right\} = - H^{-1}\left(\alpha/2, \hat{p}_A; p_A\right)$$

where p_A is considered fixed and the dependence on μ_A and σ^2_A is temporarily suppressed; a lower critical value, say c_ℓ, is defined in a similar manner. A conditional, equal tailed $(1 - \alpha) \times 100\%$ CI for T_A is then given by $\mathrm{CI}(1 - \alpha) = (\ell, u)$, where

$$u = \hat{T}_A + c_u s_{\hat{T}_A} \text{ and } \ell = \hat{T}_A + c_\ell s_{\hat{T}_A}. \qquad (3)$$

At this point the obvious practical problem is that the critical values c_u and c_ℓ depend not only on $\hat{p}_A$ but also on the unknown parameter p_A. One approach to this problem is to take a Bayesian tack and assume the parameter p_A is the realization of a random variable. Adjusting the notation to reflect the assumption that p_A is stochastic, we replace $H(x\,|\,\hat{p}_A; p_A)$ by $H(x\,|\,\hat{p}_A, p_A)$ and have that

$$\Pr\left\{\hat{\theta} \leq x \,|\, \hat{p}_A\right\} = F\left(X\,|\,\hat{p}_A\right) = \frac{1}{h(\hat{p}_A)} \int H\left(x\,|\,\hat{p}_A, p_A\right) f\left(\hat{p}_A\,|\,p_A\right) g\left(p_A\right) dp_A, \qquad (4)$$

where $h(\hat{p}_A) = \int f(\hat{p}_A\,|\,p_A) g(p_A) dp_A$ and $g(p_A)$ is the density of p_A. It should be noted that as a consequence of our sampling scheme the distribution of $n\hat{p}_A$, conditional on p_A, is Binomial (n, p_A) so that $f(\hat{p}_A\,|\,p_A)$ is known. Under the Bayesian approach, the critical values are $c^*_u \equiv c^*_u(\alpha, \hat{p}_A) = -F^{-1}(\alpha/2\,|\,\hat{p}_A)$ and $c^*_\ell \equiv c^*_\ell(\alpha, \hat{p}_A) = -F^{-1}(1 - \alpha/2\,|\,\hat{p}_A)$ so the upper and lower limits for a conditional $(1 - \alpha) \times 100\%$ CI for T_A are

$$u = \hat{T}_A + c^*_u s_{\hat{T}_A} \text{ and } \ell = \hat{T}_A + c^*_\ell s_{\hat{T}_A}. \qquad (5)$$

For the purposes of our current research, we assume that the prior distribution $g(p_A)$ is $N(\mu_{p_A}, \sigma^2_{p_A})$ with μ_{p_A} and $\sigma^2_{p_A}$ to be specified, with the understanding that $\sigma^2_{p_A}$ is sufficiently small that p_A lies between 0 and 1 with near certainty. The normality assumption is made for mathematical convenience. It also captures notions we may have of degrees of closeness to, and symmetry about, μ_{p_A}. For an empirical Bayes approach, we use $\mu_{p_A} = \hat{p}_A$; we consider several possible alternatives for $\sigma^2_{p_A}$ discussed in detail below. Our experience indicates that the normality assumption is not crucial; rather, it is primarily a matter of convenience.

2.4 Confidence Intervals Under Normal Assumptions

To proceed further we assume that within the domain A the x_i are distributed $N(\mu_A, \sigma^2_A)$. In practice, this assumption may not be met. Nonetheless, it leads to suggested modifications that will not at any rate give lower coverage of confidence intervals than the standard approach. Combining this assumption with earlier results, in particular equation (2), and ignoring lower order terms, we have

(a) $[\sqrt{n}(\hat{T}_A - T_A)/n\,|\,\hat{p}_A, p_A]$ is distributed $N(\sqrt{n}\mu_A(\hat{p}_A - p_A), \hat{p}_A\sigma^2_A)$,

(b) $\left[(n\hat{p}_A - 1)\dfrac{\hat{\sigma}^2_A}{\sigma^2_A}\,|\,\hat{p}_A, p_A\right]$ is distributed $\chi^2(n\hat{p}_A - 1)$, and

(c) the conditional random variable in (b) is stochastically independent of the conditional random variable in (a).

Consider $\hat{\theta}_1 = (\hat{T}_A - T_A)/(N\hat{\sigma}_A\sqrt{\hat{p}_A}/\sqrt{n})$, which utilizes the conditional variance of $\hat{T}_A$ as the standardizing term. It follows immediately from (a), (b) and (c) that, conditional on $(\hat{p}_A, p_A)$ the random variable $\hat{\theta}_1$ is distributed as a non-central t with $n\hat{p}_A - 1 = n_A - 1$ degrees of freedom and non-centrality parameter

$$\lambda = \sqrt{n}\gamma_A(\hat{p}_A - p_A)/\sqrt{\hat{p}_A},$$

with

$$\gamma_A = \mu_A/\sigma_A.$$

Thus, we have specified the conditional distribution function $H(\cdot \mid \hat{p}_A, p_A)$ of $\hat{\theta}_1$. As $f(\hat{p}_A \mid p_A)$ and $g(p_A)$ have been previously specified, it follows that $F(\cdot \mid \hat{p}_A)$ in (4) is well-defined although extremely cumbersome to calculate. The dependence on μ_A and σ_A^2, through γ_A, should be noted.

Although $F(\cdot \mid \hat{p}_A)$ as given above can be used to determine the critical values, they are extremely difficult to calculate. A relatively simple approach, given in the next paragraph, provides a close approximation to the critical values. We have verified the closeness of the approximation by computing the exact values for selected cases using large scale simulations.

Adoption of a locally uniform prior on p_A leads to the approximate posterior distribution $p_A \sim N(\hat{p}_A, \mathrm{var}(\hat{p}_A))$ and we could approximate $\mathrm{var}(\hat{p}_A)$ by $\hat{p}_A(1-\hat{p}_A)/n$. We adopt the slightly more flexible prior $p_A \sim N(\mu, \sigma_{p_A}^2)$, and empirically choose $\mu = \hat{p}_A$, with several possibilities for $\sigma_{p_A}^2$ that will be specified below. It follows from Appendix A that $[\lambda \mid \hat{p}_A]$ is distributed approximately as a normal with mean zero and variance $\gamma_A^2(1-\hat{p}_A)/(1+\psi_A)$, where

$$\psi_A = \hat{p}_A(1-\hat{p}_A)/n\sigma_{p_A}^2.$$

Then, from the result in Appendix B, conditional on $\hat{p}_A$,

$$\frac{\left(\hat{T}_A - T_A\right)}{\dfrac{N\hat{\sigma}_A\sqrt{\hat{p}_A}}{\sqrt{n}}\sqrt{\dfrac{\gamma_A^2(1-\hat{p}_A)}{1+\psi_A}+1}}$$

is distributed as a central t with $n_A - 1$ degrees of freedom. Let $t_{1-\alpha/2, n_A-1}$ be the $(1-\alpha/2)100\%$ percentile of this distribution. The upper confidence limit u, defined in (5), is given (approximately) by

$$u = \hat{T}_A + N\hat{\sigma}_A\sqrt{\hat{p}_A/n} \times \left(\left(\gamma_A^2\left(1-\hat{p}_A\right)+1+\psi_A\right)\Big/\left(1+\psi_A\right)\right)^{1/2} t_{1-\alpha/2, n_A-1}. \tag{6}$$

As $\hat{\sigma}_A^2$ is conditionally unbiased for σ_A^2 and $\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A$ is conditionally unbiased for μ_A^2, we use $\hat{\gamma}_A^2 = (\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A)/\hat{\sigma}_A^2$ to estimate γ_A^2. Substituting $\hat{\gamma}_A^2$ for γ_A^2 in (6) yields

$$\bar{u} \cong \hat{T}_A + \left(Ns/\sqrt{n}\right) \times \left(\left(1+\frac{\hat{p}_A\hat{\sigma}_A^2\psi_A}{s^2}\right)\Big/\left(1+\psi_A\right)\right)^{1/2} t_{1-\alpha/2, n_A-1} \tag{7}$$

where s^2 is defined in (1).

It remains to choose ψ_A. We note that $\bar{u}$ is strictly decreasing as ψ_A increases and

$$\tilde{u} \to \hat{T}_A + \frac{Ns}{\sqrt{n}} t_{1-\alpha/2, n_A-1} = \tilde{u} \text{ as } \psi_A \text{ becomes small,}$$

$$\tilde{u} = \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\frac{1+\hat{p}_A\hat{\sigma}_A^2/s^2}{2}\right)^{1/2} t_{1-\alpha/2, n_A-1} = \bar{u}_2 \text{ for } \psi_A = 1,$$

and

$$\tilde{u} \to \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\frac{\sqrt{\hat{p}_A}\hat{\sigma}_A}{s}\right) t_{1-\alpha/2, n_A-1} = \tilde{u}_3$$

$$\text{as } \psi_A \text{ becomes large.} \tag{8}$$

In each case the lower critical value can be dealt with in an analogous manner resulting in three competing confidence intervals; namely, $\mathrm{CI}_i(1-\alpha) = (\tilde{\ell}_i, \bar{u}_i)$, $i = 1, 2, 3$, with $\tilde{\ell}_i$ defined similarly to $\tilde{u}_i$ in (8) with $t_{1-\alpha/2, n_A-1}$ replaced by $t_{\alpha/2, n_A-1}$. The competing confidence intervals are labeled in order of decreasing length.

The first case is equivalent to assuming that $\sigma_{p_A}^2$ is large relative to $\mathrm{var}(\hat{p}_A)$ and leads to using the usual unconditional variance but with degrees of freedom equal to $n_A - 1$. In most practical problems this seems reasonable since $\sigma_{p_A}^2$ is an unknown constant and $\mathrm{var}(\hat{p}_A)$ is $0(p_A/n)$. The second interval corresponds to adoption of a normal prior as noted above, with $\sigma_{p_A}^2 = \hat{p}_A(1-\hat{p}_A)/n$. The last confidence interval is based on the assumption that p_A is essentially degenerate at $\hat{p}_A$.

2.5 Empirical Study for SRS

We compared the several confidence intervals of Section 2.4 in a small empirical study, using artificial populations, for which the domain variable was normal. In all cases the population size N was 1,000, and the sample size n was 100 or 300. The parameters p_A and γ_A varied from population to population. Letting M_2 be the number of runs with $n_A \geq 2$, we allowed the run size M to vary to give $M_2 = 10{,}000$. Table 1 gives coverage results. CI_0 represents the confidence interval based on the standard normal methodology. The results for CI_2 closely approximated the results for CI_1 and are excluded. The value of M is included to indicate how many trials fell into the "insufficient information" pile, at a given setting of the parameters. Several conclusions seem warranted:

1. Standard confidence intervals using the usual variance estimate and normal quantiles can give low coverage. This occurs for several values of p_A when $\gamma_A = 1/2$ or $\gamma_A = 2$, however, the under-coverage is not too severe if the domain variable is normal. The case where

$\gamma_A = 2$ or takes even larger values is probably more likely in practice. Thus if the domain variable is normal, the use of standard confidence intervals under simple random sampling case is not particularly worrisome.

2. The strictly conditional intervals (*i.e.*, CI_3) using the conditional variance can give abominable coverage, when γ_A is large. That is, confidence intervals based on "large" values of ψ_A gave very poor results.
3. The use of the standard variance estimate but replacing the standard normal quantile with a t-quantile having degrees of freedom based on the number of sample units in the domain (*i.e.*, CI_1) gives approximately nominal or conservative coverage regardless of the value of γ_A.

Table 1
Coverage of 95% Confidence Intervals for Domain Total
for Artificial Populations with
Domain Variate Normally Distributed*

			Coverage		
P_A	n	M	CI_0	CI_1	CI_3
			$y = 1/2$		
.01	100	38774	100.0	100.0	91.2
	300	11773	98.3	100.0	83.2
.02	100	16327	91.1	99.4	95.0
	300	10078	88.6	95.5	93.9
.05	100	10303	88.7	97.8	93.5
	300	10000	92.3	94.4	92.5
.10	100	10001	90.9	94.8	92.5
	300	10000	94.0	95.0	92.3
			$y = 2$		
.01	100	37749	99.9	100.0	83.5
	300	11740	94.4	100.0	89.1
.02	100	16348	99.0	100.0	88.4
	300	10075	91.4	98.9	74.7
.05	100	10312	90.5	99.5	77.6
	300	10000	93.8	95.8	66.6
.10	100	10000	91.7	96.5	67.9
	300	10000	94.0	95.2	65.0

* See Equation (8) and accompanying text for definition of CI_1 and CI_3. CI_0 is the standard normal confidence interval.

As a minor observation on the results, we note the counter-intuitive increases in coverage for smaller p_A and n. We believe this is due to the fact that, at very small values of p_A and n, $\hat{p}_A$ is constrained to be positive, and so cannot deviate much below p_A. Were intervals calculable for $n_A = 0$, there would be a serious drop in coverage in these cases. Note that the coverage rises unexpectedly only where M is large.

3. THE CASE OF STRATIFIED RANDOM SAMPLING

3.1 Definitions and Notation

Assume there are K strata and, where appropriate, terms previously defined have corresponding stratum level definitions. For example, n_k is the sample size and n_{Ak} is the number of sample elements in A for the k-th stratum. Thus, a natural estimator for the domain total $T_A = \sum_{k=1}^K \sum_{i \in A} x_{ki} = \sum_{k=1}^K N_k \hat{p}_{Ak} \mu_{Ak}$ is

$$\hat{T}_A = \sum_{k \in B_1} \hat{T}_{Ak} = \sum_{k \in B_1} N_k \hat{p}_{Ak} \hat{\mu}_{Ak},$$

where $\hat{p}_{Ak} = n_{Ak}/n_k$, $\hat{\mu}_{Ak} = \sum_{i=1}^{n_{Ak}} x_{ki}/n_{Ak}$ and $B_1 = \{k \mid n_{Ak} \geq 1$ and $1 \leq k \leq K\}$. As $\hat{p}_{Ak} = 0$ for $k \notin B_1$, it is straightforward to verify that

$$E\left[(\hat{T}_A - T_A) \mid \hat{\boldsymbol{p}}_A, \boldsymbol{p}_A\right] = \sum_{k=1}^K N_k(\hat{p}_{Ak} - p_{Ak})\mu_{Ak} \equiv \tilde{\mu}_A \quad (9)$$

and

$$\mathrm{var}\left[(\hat{T}_A - T_A) \mid \hat{\boldsymbol{p}}_A, \boldsymbol{p}_A\right] = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak}^2 \sigma_{Ak}^2 / n_{Ak} =$$

$$\sum_{k \in B_1} N_k^2 \hat{p}_{Ak}^2 \sigma_{Ak}^2 / n_k \equiv \tilde{\sigma}_A^2,$$

where $\hat{\boldsymbol{p}}_A = [\hat{p}_{A1} \hat{p}_{A2} \cdots \hat{p}_{AK}]$, $\boldsymbol{p}_A = [p_{A1} p_{A2} \cdots p_{AK}]$. Thus, as in the simple random sampling case, there is a conditional bias $\tilde{\mu}_A$, which needs to be taken into account.

3.2 A Methodology for Confidence Intervals

The general methodology for confidence intervals of Section 2.3 for simple random sampling holds here as well. One need only reinterpret scalars as vectors; for example, replace $\hat{p}_A$ by $\hat{\boldsymbol{p}}_A = (\hat{p}_{A1}, \ldots, \hat{p}_{AK})'$. In particular, $H(x \mid \hat{\boldsymbol{p}}_A, \boldsymbol{p}_A) = \Pr\{\hat{\theta} \leq x \mid \hat{\boldsymbol{p}}_A, \boldsymbol{p}_A\}$ will be the conditional distribution function of $\hat{\theta} = (\hat{T}_A - T_A)/\hat{\sigma}_A$, where $\hat{\sigma}_A$ is a re-scaling factor to be specified.

Let $B_2 = \{k \mid n_{Ak} \geq 2$ and $1 \leq k \leq K\}$ and, for $k \in B_2$, define $\hat{\sigma}_{Ak}^2 = \sum_{i=1}^{n_{Ak}} (x_{ki} - \hat{\mu}_{Ak})^2/(n_{Ak} - 1)$. Under normality, $(n_{Ak} - 1)\hat{\sigma}_{Ak}^2/\sigma_{Ak}^2 \sim \chi^2(n_{Ak} - 1)$, so if $\{d_k \mid k \in B_2\}$ are non-negative constants with $\sum_{k \in B_2} d_k > 0$, then by the usual Satterthwaite (1946) two moment approximation, the conditional random variable

$$\left[(1/c) \sum_{k \in B_2} d_k (n_{Ak} - 1)(\hat{\sigma}_{Ak}^2/\sigma_{Ak}^2) \mid \hat{\boldsymbol{p}}_A, \boldsymbol{p}_A\right]$$

is distributed approximately as a $\chi^2(\nu)$, where

$$c = \sum_{k \in B_2} d_k^2 (n_{Ak} - 1) \Big/ \sum_{k \in B_2} d_k (n_{Ak} - 1)$$

and

$$\nu = \left(\sum_{k \in B_2} d_k (n_{Ak} - 1)\right)^2 \Big/ \sum_{k \in B_2} d_k^2 (n_{Ak} - 1).$$

This suggests that we restrict our attention to expressions of the general form

$$\hat{\sigma}_A^2 = \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$$

with choice of the d_k to be specified. Note that when $B_1 = B_2$ and $d_k = N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k (n_{Ak} - 1)$, $\hat{\sigma}_A^2 = \hat{\tilde{\sigma}}_A^2 \equiv \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$ is an unbiased estimator for the conditional variance $\tilde{\sigma}_A^2$. However, as in the simple random sampling case, this estimator will tend to be too small. We use the more general expression to develop a family of t-statistics when we "uncondition" on p_A. Each of these will involve unknown parameters, and, as in the simple random sampling case (transition of equation (6) to equation (7)), estimation of these unknowns will be necessary. Thus the net result will be several rival "near t-statistics" which we may then compare empirically.

Because the samples are selected independently from each stratum we have $f(\hat{p}_A | p_A) = \Pi_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak})$ and, as a consequence of our within stratum sampling scheme, $n_k \hat{p}_{Ak}$ has a binomial distribution $B(n_k, p_{Ak})$. We assume that the $\{p_{Ak} | 1 \le k \le K\}$ are jointly independent so $g(p_A) = \Pi_{k=1}^K g_k(p_{Ak})$ which implies

$$f(\hat{p}_A | p_A) g(p_A) = \prod_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak})$$

and

$$h(\hat{p}_A) = \prod_{k=1}^K \int f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak}) dp_{Ak}.$$

In what follows, we assume that the prior distribution of p_{Ak} is $N(\mu_{p_{Ak}}, \sigma_{p_{Ak}}^2)$ and for the empirical Bayes approach, we use $\mu_{p_{Ak}} = \hat{p}_{Ak}$ and, analogously to the case of simple random sampling, we define

$$\psi_{Ak} = \hat{p}_{Ak}(1 - \hat{p}_{Ak}) / n_k \sigma_{p_{Ak}}^2.$$

It is straightforward to extend the result in Appendix A to the case of stratified random sampling and it then follows that, for $\tilde{\mu}_A$ defined by (9), $[\tilde{\mu}_A / \tilde{\sigma}_A | \hat{p}_A]$ is distributed $N(0, \text{var}(\bar{\mu}_A | \hat{p}_A) / \tilde{\sigma}_A^2)$, where $\text{var}(\bar{\mu}_A | \hat{p}_A) = \sum_{k \in B_1} N_k^2 \mu_{Ak}^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) / n_k (1 + \psi_{Ak})$. Using the result in Appendix B, it follows that, conditional on $\hat{p}_A$, the random variable

$$\hat{\theta} = \frac{(\hat{T}_A - T_A) \Big/ \sqrt{\text{var}(\tilde{\mu}_A | \hat{p}_A) + \tilde{\sigma}_A^2}}{\sqrt{\hat{\sigma}_A^2 / cv}} =$$

$$\frac{(\hat{T}_A - T_A) \Big/ \sqrt{\text{var}(\bar{\mu}_A | \hat{p}_A) + \bar{\sigma}_A^2}}{\sqrt{\sum_{k \in B_1} d_k (n_{Ak} - 1)(\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2) \Big/ \sum_{k \in B_1} d_k (n_{Ak} - 1)}}$$

is distributed approximately as a central t with v degrees of freedom.

Letting $\Theta = \text{var}(\tilde{\mu}_A | \hat{p}_A) + \tilde{\sigma}_A^2$, with

$$\gamma_{Ak}^2 = \mu_{Ak}^2 / \sigma_{Ak}^2$$

and assuming the ψ_{Ak} are near zero we have

$$\Theta = \sum_{k \in B_1} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1).$$

Thus, the upper bound on the CI would be (approximately)

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1)(\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)}}{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1)}} \Theta^{1/2} t_v, \tag{10}$$

where t_v stands for the critical values of the t_v distribution. Unfortunately the bound depends not only on our choice of the d_k, but also on the unknown parameters μ_{Ak} and σ_{Ak}^2.

It is not hard to show that $v \le \sum_{k \in B_2} (n_{Ak} - 1) \equiv v_{\max}$ and, if we set $d_k = 1$ (or any constant for that matter) then $v = v_{\max}$. We refer to $v_{\max}$ specifically as the unweighted degrees of freedom. In this case the upper bound on the CI would be

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1)(\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)}}{\sqrt{\sum_{k \in B_2} (n_{Ak} - 1)}} \Theta^{1/2} t_{v_{\max}}.$$

Another approach is to attempt to finesse the problem of estimating Θ (at least when $B_1 = B_2$) by a judicious choice of the d_k. To that end let us assume that $B_1 = B_2$ and let

$$d_k = \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k (n_{A_k} - 1)} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)$$

so that $\sum_{k \in B_2} d_k (n_{Ak} - 1) = \Theta$ and Θ cancels out in (10). We then have

$$u = \hat{T}_A + \sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \hat{\sigma}_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1) t_{v_1}},$$

where v_1 is the degrees of freedom associated with this second choice of the d_k. More generally (*i.e.*, when $B_1 \ne B_2$), we have

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \hat{\sigma}_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)}}{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)}} \Theta^{1/2} t_{v_1}.$$

In any event, we are still faced with the problem of estimating the population parameters and we have the additional problem of estimating the degrees of freedom.

A third possibility, which we have already mentioned, is to let $d_k = N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k(n_{Ak} - 1)$ so that when $B_1 = B_2$, $\hat{\sigma}_A^2 = \hat{\tilde{\sigma}}_A^2 \equiv \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$ is a conditionally unbiased estimator for $\bar{\sigma}_A^2$. In this case we have

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} N_k^2 \hat{p}_k \hat{\sigma}_{Ak}^2 / n_k}}{\sqrt{\sum_{k \in B_2} N_k^2 \hat{p}_k \sigma_{Ak}^2 / n_k}} \Theta^{1/2} t_{\nu_2},$$

where ν_2 is the degrees of freedom associated with this third choice of the d_k. As in the second case, we are faced with the problem of estimating the population parameters and the degrees of freedom.

Now, it should be noted that if we estimate σ_{Ak}^2 with $\hat{\sigma}_{Ak}^2$ for $k \in B_2$ and let $\hat{\Theta}$ be a yet to be specified estimator of Θ then the (estimated) upper bounds above are $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{\nu_{\max}}$, $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{\hat{\nu}_1}$ and $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{\hat{\nu}_2}$ respectively. The degrees of freedom are estimated by substituting estimates of the population parameters into the two respective choices of the d_k. Both $\hat{\nu}_1$ and $\hat{\nu}_2$ are smaller than $\nu_{\max}$, so, for any realized value of $\hat{\Theta}$, the confidence interval using $\nu_{\max}$ will be the shortest. There is no general relationship between the sizes of $\hat{\nu}_1$ and $\hat{\nu}_2$. Empirical evidence indicates that there is little to choose between the second and third approach.

Addressing the problem of estimating Θ, we can write

$$\Theta = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{Ak} \left(\mu_{Ak}^2 (1 - \hat{p}_{Ak}) + \sigma_{Ak}^2 \right) \Big/ n_k +$$

$$\sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \left(\mu_{Ak}^2 (1 - \hat{p}_{Ak}) + \sigma_{Ak}^2 \right) \Big/ n_k.$$

For $k \in B_1 - B_2$ the estimator $\hat{\sigma}_{Ak}^2$ is not defined, however, it is straightforward to verify that $(1 - \hat{p}_{Ak}) E[\hat{\mu}_{Ak}^2 | n_{Ak}] \leq \sigma_{Ak}^2 + \mu_{Ak}^2 (1 - \hat{p}_{Ak}) \leq E[\hat{\mu}_{Ak}^2 | n_{Ak}]$. It follows that

$$s_a^2 = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 / n_k +$$

$$\sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 + 1/n_k - 1/n_{Ak}) / n_k$$

will tend to underestimate Θ, and

$$s_b^2 = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{Ak} \hat{\mu}_{Ak} \Big/ n_k + \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 \Big/ n_k +$$

$$\sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 + 1/n_k - 1/n_{Ak}) / n_k$$

will tend to overestimate Θ. Clearly, $s_a^2 \leq s_b^2$ with equality only when $B_1 = B_2$.

It can also be verified that in the case of stratified sampling, the standard variance estimator for estimated population totals is

$$s_{\text{std}}^2 = \sum_{k \in B_1} N_k^2 s_k^2 \Big/ n_k = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 \Big/ (n_k - 1)$$

$$+ \sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 - 1/n_{Ak}) / (n_k - 1).$$

This looks like a satisfactory estimator of Θ, if the n_k are not small.

These results imply that CIs of the form $(\hat{T}_A \pm s_b t_{1-\alpha/2, \hat{\nu}_1})$ will provide the highest level of coverage; but CIs of the form $(\hat{T}_A \pm s_{\text{std}} t_{1-\alpha/2, \nu_{\max}})$ and even perhaps $(\hat{T}_A \pm s_{\text{std}} t_{1-\alpha/2, \hat{\nu}_1})$ have obvious computational advantages. Several of these competing forms of CI are evaluated empirically in Section 3.3. These results can easily be extended to ratio estimators by the standard linearization approach.

3.3 Empirical Investigation for Stratified Random Sampling: the BLS Wage Data

With a view to improving estimation of precision on wage data produced by the U.S. Bureau of Labor Statistics, we investigated coverage and interval length in two simulation studies on populations constructed from a test sample of the Occupational Compensation Survey Program (OCSP) conducted in 1991. The OCSP consisted of establishment surveys in several metropolitan areas, aimed at estimating wages levels for a select group of occupations. The surveys were carried out by stratified simple random sampling, with establishments stratified by employment size and industrial classification.

One population (the "Small Population") took the test sample itself as the population, with six non-certainty strata, and one certainty stratum of 12 establishments. Five hundred stratified random samples were taken from this population at sizes $n = 36$ and 60, corresponding to the choices $n_k = 4$ and $n_k = 8$, reflecting relative sample sizes of sampling from the original population. The second population (the "Large Population") was constructed by expanding the sample data through replication (by simple random sampling with replacement, within each Small Population stratum) of establishments to achieve a population the size of the original population; again there were six noncertainty and one certainty strata; for each stratum sample sizes were the same as in the actual sample. Domains are defined by the different occupations of interest; only a fraction of establishments have workers in a particular occupation, and lie in the corresponding domain. Table 2 gives the number of establishments having workers in the selected occupations for the small population.

In both cases sampling was without replacement, so finite population correction factors were included (as appropriate) in the construction of the CIs. Also, the study was limited to a concern with 95% coverage.

Table 2
Number of Establishments in Given Domain (Occupation), by Stratum for Small Population

	stratum							
Occupation	1	2	3	4	5	6	7	total
4021	0	4	11	10	8	10	7	50
1141	0	3	11	7	11	9	7	48
1122	0	3	8	13	14	12	6	56
3180	10	11	5	25	20	4	5	80
2911	0	3	14	2	13	17	7	56
1142	2	8	15	9	15	19	9	77
1180	17	20	5	61	31	3	1	138
1403	12	16	22	28	25	27	9	139
All Estabs	35	35	33	136	66	36	12	353

Small Population: Table 3 gives coverage and median relative interval length for total wages, at two sample sizes $n_k = 4$ and $n_k = 8$, for 8 occupations, and three methods of confidence interval construction: the standard variance estimator, s^2_{std}, with the standard normal z-quantile, the unweighted degrees of freedom $\nu_{\max}$, and the weighted degrees of freedom ν_1. Occupations are ordered by increasing values of the average value, over runs, of the unweighted degrees of freedom. We note:

1) Almost universally, coverage using the standard variance estimator and the standard normal quantiles (infinite *df*) is poor.
2) Coverage for the other interval types is far more satisfactory. In general, the coverage is near the nominal 95%, or slightly conservative, for weighted degrees of freedom; as expected, intervals based on unweighted degrees of freedom tend to yield coverage a few points below those based on weighted degrees of freedom.
3) Two occupations (1122, 4021) yield seriously low coverage for totals even with the improved procedures. Investigation of these particular occupations suggests a strong violation of the normality assumption. In 4021, for example, two units in stratum 5 have a number of workers, and hence total wages, an order of magnitude higher than the other establishments in this stratum and indeed in the population. Furthermore, the wage rate of these two outliers is markedly lower than the great bulk of establishments: with just these two excluded from the population, the overall population average wage would be \$9.68/hour; with them in, it is \$8.28. Since there are 66 establishments in stratum 5, it is easy for these two establishments to escape being in a sample of size 8; the consequence is a serious overestimate of the mean wage or underestimate of total wage. At the same time, wages for the establishments that are in the sample are relatively homogeneous, so the variance estimate will tend to be too low. The presence of several smaller establishments in the domain contribute to enlarging the degrees of freedom, and so the t-adjustment is unable to compensate fully. It is hard to see how to guard against such a problem short of having prior information, and allotting such outliers to a certainty stratum. Even so, the adjusted intervals are a significant improvement on the naïve normal distribution based interval.

Interval lengths are taken relative to $2 \times z_{.975} \approx 4$ times the root mean square error of $\hat{T}_A$ calculated over runs. We report the median of these standardized lengths (across runs). When the distribution of $\hat{T}_A$ is actually normal, the median length is close to 1.

Table 3
Estimated degrees of freedom, coverage, and relative median length of CIs for total wages of workers in occupation, for the small population

	Four Sample Establishments Per Stratum								Eight Sample Establishments Per Stratum							
Occupation	4021	1141	1122	3180	2911	1142	1180	1403	1141	4021	1122	3180	2911	1142	1180	1403
$df = \nu_{\max}$	1.5	1.6	1.6	2.0	2.3	2.8	4.3	6.1	3.7	3.8	3.9	5.6	6.0	8.0	12.3	16.6
$df = \hat{\nu}_1$	1.3	1.3	1.4	1.5	1.7	1.9	2.3	3.5	2.0	2.3	2.3	3.1	3.5	4.3	5.4	9.7
								Coverage								
$\hat{T}_A \pm s_{\text{std}} z$	.47	.69	.51	.75	.73	.85	.89	.87	.74	.49	.65	.79	.78	.86	.88	.92
$\hat{T}_A \pm s_{\text{std}} t_{\nu_{\max}}$	.89	.92	.93	.99	.95	.96	.97	.92	.87	.65	.75	.89	.86	.90	.90	.94
$\hat{T}_A \pm s_{\text{std}} t_{\hat{\nu}_1}$	.92	.93	.95	.99	.96	.96	.98	.95	.91	.74	.80	.94	.89	.95	.96	.96
								Median Relative Length								
$\hat{T}_A \pm s_{\text{std}} z$	0.53	0.75	0.59	0.70	0.74	0.85	0.90	0.88	0.87	0.63	0.66	0.80	0.83	0.88	0.92	0.96
$\hat{T}_A \pm s_{\text{std}} t_{\nu_{\max}}$	2.65	3.67	2.80	2.60	2.20	1.98	1.50	1.14	1.63	1.09	1.13	1.10	1.10	1.06	1.02	1.04
$\hat{T}_A \pm s_{\text{std}} t_{\hat{\nu}_1}$	3.30	4.32	3.19	3.40	3.08	3.06	2.70	1.58	3.08	2.40	2.38	2.00	1.74	1.38	1.38	1.13

4) The relative interval length of the standard interval tends to be too small, that is, it tends to be less than 1.
5) Interval length among the other variance-degrees of freedom combinations is largest for s_{std}^2 with $\hat{\nu}_1$, and smallest for s_{std}^2 with $\nu_{\max}$. These differences can be appreciable; there is a tradeoff between coverage and interval size.
6) For a given interval type, the relative interval length tends to 1 as $\nu_{\max}$ increases. The conclusions from a study of mean wages are similar.

Large Population: Table 4 gives coverage and interval length for total wages for five interval types, and a wider range of occupations, ordered by average $\nu_{\max}$. The interval types include the three used previously for the small population. The two new intervals utilize the weighted degrees of freedom together with s_a and s_b respectively. Results are based on 5,000 runs.

1) The results are consistent with those for the Small Population, in terms of the relative coverage and interval sizes of the several interval types. The standard normal is unsatisfactory for many occupations.
2) The coverage for intervals using the weighted degrees of freedom, $\hat{\nu}_1$, is less than 90% for only a small fraction of cases.
3) There can be marked differences in interval length for the different interval types; however, all ratios of interval length to 4 × root mean square error tend to 1, as $\nu_{\max}$ gets large.
4) Little difference results from using s_a, s_b, or s_{std} with $t_{\hat{\nu}_1}$.

Again, the results for mean wages, while differing in detail, lead to the same overall conclusions, and are omitted.

4. SUMMARY AND CONCLUSIONS

From our theoretical investigation and simulation work, we draw the following conclusions:

1. Standard 95% confidence intervals for domain means or totals, when based on the standard normal distribution and standard methods of variance estimation, tend to yield less than actual 95% coverage. The extent of the deviation will vary with domain (occupation in the wage study), but can be quite considerable even when the sample size is large.
2. New nonstandard methods offer a sharp improvement, giving intervals with better coverage, typically at or close to the nominal 95% coverage. These intervals tend to be longer than the standard intervals. The increase in length will vary with domain, and will depend on the particular method for CI construction that is adopted.

Table 4

Estimated degrees of freedom, coverage, and relative median length of CIs for total wages of workers in occupation, for the large population

	Occupation																				
	1718	1604	1802	1716	2911	2052	1332	1141	4021	1232	2853	3020	1122	1142	1714	1514	3180	4030	1063	1403	1180
$df = \nu_{\max}$	2.97	3.45	4.44	11.9	12.4	13.1	15.3	16.9	16.8	17.3	20.6	24.9	28.0	28.6	29.1	34.8	41.5	59.9	77.6	77.9	128
$df = \hat{\nu}_1$	2.67	2.34	2.35	5.97	5.90	4.25	11.4	9.00	6.32	15.5	13.5	10.4	15.2	9.67	15.3	18.0	25.2	14.3	27.4	28.5	90.0
Coverage																					
$\hat{T}_A \pm s_{std} z$	.89	.60	.85	.87	.87	.89	.93	.93	.89	.92	.92	.92	.88	.89	.85	.93	.92	.81	.94	.94	.94
$\hat{T}_A \pm s_{std} t_{\nu_{\max}}$	.96	.83	.94	.89	.88	.91	.95	.95	.91	.94	.94	.93	.88	.90	.86	.93	.92	.81	.95	.94	.95
$\hat{T}_A \pm s_a t_{\hat{\nu}_1}$	.97	.88	.94	.91	.89	.97	.96	.96	.91	.94	.94	.95	.89	.91	.86	.94	.93	.83	.95	.94	.95
$\hat{T}_A \pm s_{std} t_{\hat{\nu}_1}$	.97	.89	.94	.92	.90	.97	.96	.91	.94	.94	.95	.89	.89	.91	.86	.94	.93	.83	.95	.95	.95
$\hat{T}_A \pm s_b t_{\hat{\nu}_1}$	.97	.89	.97	.92	.90	.97	.96	.96	.91	.95	.94	.95	.89	.91	.87	.95	.93	.83	.95	.94	.95
Median Relative Length																					
$\hat{T}_A \pm s_{std} z$	0.99	0.78	0.92	0.97	0.95	0.96	0.99	0.98	0.96	0.97	0.98	.98	0.95	0.96	0.93	0.98	1.00	0.91	1.00	1.00	1.01
$\hat{T}_A \pm s_{std} t_{\nu_{\max}}$	2.14	1.47	1.40	1.08	1.06	1.06	1.08	1.06	1.04	1.04	1.04	1.03	0.99	1.00	0.98	1.01	1.03	0.93	1.01	1.01	1.02
$\hat{T}_A \pm s_a t_{\hat{\nu}_1}$	2.32	2.24	2.46	1.37	1.37	1.59	1.12	1.15	1.34	1.05	1.11	1.16	1.04	1.19	1.04	1.04	1.05	1.07	1.09	1.04	1.02
$\hat{T}_A \pm s_{std} t_{\hat{\nu}_1}$	2.34	2.27	2.48	1.37	1.39	1.60	1.13	1.18	1.34	1.05	1.13	1.18	1.04	1.20	1.04	1.04	1.06	1.07	1.10	1.05	1.02
$\hat{T}_A \pm s_b t_{\hat{\nu}_1}$	2.47	2.33	2.79	1.39	1.38	1.61	1.14	1.20	1.35	1.07	1.13	1.18	1.04	1.19	1.05	1.05	1.06	1.07	1.10	1.04	1.02

For domains which yield large samples, there will be little difference from standard intervals.

3. The instances where coverage fell below nominal, even using the t-adjusted intervals, may be ascribed to severe violation of the normality assumption for the domain data. Thus the t-adjustment is not a cure-all. Nonetheless, even in such cases there is a good deal of improvement in coverage over the use of the standard normal interval.

4. The key idea behind these intervals is to condition on the amount of information on the particular occupation, which, roughly speaking, is measured in terms of the number of units in the sample that belong to the domain. The fraction of such units within each stratum is unknown, and to handle this fact we put a prior distribution on this unknown, reflective of the degree of our ignorance of it, an idea we borrow from the Bayesians. However, in the final analysis, it is the realized coverage probabilities that determine the merit of the approach.

5. The principal effect of these ideas is the abandonment, for purposes of CI construction, of the standard normal quantiles (± 1.96 for 95% coverage). These are replaced by quantiles from the Student's t-distribution, with degrees of freedom determined from the sample and varying with domain. If because of publication requirements or for other reasons, there is need to report standard deviations rather than confidence intervals, then we recommend reporting an effective standard deviation given by the length of the proposed t-based 95% confidence interval divided by twice 1.96.

6. The standard estimate of variance seems acceptable for estimating the variance, when accompanying the new t-quantile. In most instances this combination should be quite satisfactory, so that the only change from standard methodology will be the introduction of adjusted degrees of freedom. However, in some instances, the alternative standard deviations may improve coverage or reduce the length of confidence intervals.

7. An open question concerns what degree and type of collapsing of strata (if any) should be used in the estimation of variances and of the degrees of freedom for the purpose of confidence interval construction. In general, there will be a tradeoff: as strata are reduced in number, the estimate of variance will tend to increase, but so will the degrees of freedom (reducing the size of $t_{v_{\max}}$ or $t_{\hat{\nu}_1}$.) The answer to this question may be population specific, and experience from past surveys useful.

ACKNOWLEDGEMENTS

The authors wish to thank the Associate Editor and the three referees for their many helpful suggestions. The research of S. Wang was supported in part by an ASA/NSF/BLS research fellowship, and grants from the National Security Agency (MDA904-96-1-0029), the National Science Foundation (DMS-9504589) and the Texas A&M University Scholarly and Creative Activities Program (95-59). The views are those of the authors and do not necessarily reflect the U.S. Bureau of Labor Statistics policy.

APPENDIX A

From the discussion in Section 2.2 we know that $n\hat{p}_A$ has a binomial distribution $Bin(n, p_A)$, hence, for $\hat{p}_A = 0, 1/n, 2/n, ..., 1$,

$$f(\hat{p}_A | p_A) = \frac{\Gamma(n+1) \qquad \Gamma(n+2)}{\Gamma(n+2)\Gamma(n\hat{p}_A+1)\Gamma(n(1-\hat{p}_A)+1)} \times$$

$$p_A^{(n\hat{p}_A+1)-1}(1-p_A)^{(N(1-\hat{p}_A)+1)-1} = k_{\hat{p}_A}(p_A)/(n+1).$$

For each (fixed) value of $\hat{p}_A$, the function $k_{\hat{p}_A}(p_A)$ is the pdf of a Beta distribution with parameters $\omega_1 = n\hat{p}_A + 1$ and $\omega_2 = n(1-\hat{p}_A) + 1$. As both ω_1 and ω_2 will be larger than unity with high probability (at least in most real world situations), it is reasonable to approximate $k_{\hat{p}_A}(p_A)$ with a normal pdf having equivalent mean and variance, which are approximately $\hat{p}_A$ and $\hat{p}_A(1-\hat{p}_A)/n$ respectively.

Assuming that $p_A \sim N(\mu, \sigma^2)$, it follows that the posterior distribution is

$$h(p_A | \hat{p}_A) = f(\hat{p}_A | p_A)g(p_A) \Big/ \int_0^1 f(\hat{p}_A | p_A)g(p_A)dp_A \cong ce^{-\frac{1}{2}\left(\frac{(p_A-\hat{p}_A)^2}{\hat{p}_A(1-\hat{p}_A)/n} + \frac{(p_A-\mu)^2}{\sigma^2}\right)},$$

where c is the normalizing constant.

Under the "empirical Bayes" assumption that $\mu = \hat{p}_A$ and $\sigma^2 = \hat{p}_A(1-\hat{p}_A)/n$ we have

$$h(p_A | \hat{p}_A) \cong \frac{1}{\sqrt{2\pi}\sqrt{\hat{p}_A(1-\hat{p}_A)/2n}} e^{-\frac{1}{2}\left(\frac{(p_A-\hat{p}_A)^2}{\hat{p}_A(1-\hat{p}_A)/2n}\right)}.$$

If we drop the specific assumption regarding σ^2, and let $\psi = (\hat{p}_A(1-\hat{p}_A)/n)/\sigma^2$ then $[p_A | \hat{p}_A] \sim N(\hat{p}_A, \hat{p}_A(1-\hat{p}_A)/(1+\psi)n)$.

APPENDIX B

Result: Assume W is distributed $N(0, c^2)$ and, conditional on $W = w$, the random variable T is distributed as a non-central t with ν degrees of freedom and non- centrality parameter w. Then, the unconditional distribution of $T/\sqrt{c^2 + 1}$ is central t with ν degrees of freedom.

Proof: First notice that T can be written as $T = (X + W)/\sqrt{S^2/\nu}$, where X is distributed as $N(0, 1)$, S^2 is distributed as χ^2_ν, and X, W, and S^2 are mutually independent. Therefore, $X' = (X + W)/\sqrt{1 + c^2}$ is distributed as $N(0, 1)$. As X' and S^2 are independent, it follows by definition that $T' = T/\sqrt{1 + c^2} = X'/\sqrt{S^2/\nu}$ is distributed as t_ν.

REFERENCES

DORFMAN, A., and VALLIANT, R. (1993). Quantile variance estimators in complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 866-871.

JOHNSON, E.G., and RUST, K.F. (1993). Effective Degrees of Freedom for Variance Estimates from a Complex Sample Survey. Paper presented at the 1993 Joint Statistical Meetings, San Francisco.

KOTT, P.S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology*, 20, 159-164.

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

SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 69-77
Statistics Canada

On Regression Estimation of Finite Population Means

GIORGIO E. MONTANARI[1]

ABSTRACT

This paper examines the main properties of the generalized regression estimator of a finite population mean and those of the regression estimator obtained from the optimal difference estimator. Given that the latter can be more efficient than the former, conditions allowing this to happen are established, and a criterion for choosing between the two types of regression estimators follows. A simulation study illustrates their finite sample performances.

KEY WORDS: Generalized regression estimator; Difference estimator; Auxiliary information.

1. INTRODUCTION

Regression estimation is an effective technique for estimating survey variable finite population means or totals when the population means or totals of a set of auxiliary variables are known. The problem can be stated as follows. Consider a finite population $\mathcal{P} = \{a_1, a_2, ..., a_N\}$ consisting of N units labelled 1, 2, ..., N. Let Y_i be the value of unit a_i of a survey variable y whose population mean $\bar{Y} = \sum_1^N Y_i/N$ has to be estimated by means of a sample drawn from $\mathcal{P}$. To this end let us suppose that the population mean $\bar{\mathbf{X}} = \sum_1^N \mathbf{x}_i/N$ of a q-dimensional auxiliary variable vector, having $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{qi})'$ as its value for unit a_i, is known, for example from administrative registers or a census. The entries of $\mathbf{x}_i$ can be quantitative as well as indicator variables denoting the membership of the unit to given subpopulations. Let s be the set of sample unit labels obtained from a sampling design having first order inclusion probabilities π_i, $i = 1, 2, ..., N$, strictly positive. Then, a regression estimator can be written as follows

$$\hat{\bar{Y}}_r = \hat{\bar{Y}} + (\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}})' \hat{\boldsymbol{\beta}}, \tag{1}$$

where $\hat{\bar{Y}} = \sum_{i\in s} Y_i/N\pi_i$ and $\hat{\bar{\mathbf{X}}} = \sum_{i\in s} \mathbf{x}_i/N\pi_i$ are the Horvitz-Thompson unbiased estimators of $\bar{Y}$ and $\bar{\mathbf{X}}$, respectively, and $\hat{\boldsymbol{\beta}}$ is a vector of regression coefficients, given by some function of sample data $\{(Y_i, \mathbf{x}_i'), i\in s\}$. Briefly, $\hat{\bar{Y}}_r$ is obtained by adding to the unbiased estimator $\hat{\bar{Y}}$ terms proportional to the difference between the true means of the auxiliary variables, $\bar{X}_k = \sum_1^N x_{ki}/N$, $k = 1, 2, ..., q$, and the corresponding estimates $\hat{\bar{X}}_k = \sum_{i\in s} x_{ki}/N\pi_i$.

This paper discusses the two chief methods of constructing the vector $\hat{\boldsymbol{\beta}}$ and the properties of the corresponding regression estimators. A criterion based on a first order approximation analysis is then given for selecting one of the two alternatives. Finally, the results of two empirical studies, carried out to explore the finite sample performances of the examined estimators, are reported. All unsubscripted expectations and variances are taken with respect to a sample design. When calculations are made with respect to a model, a subscript m will be used.

2. MAIN PROPERTIES OF THE REGRESSION ESTIMATOR

Mild restrictions on the second order inclusion probabilities of the sampling design and on the limiting population moments of Y_i and $\mathbf{x}_i$ are sufficient to ensure that the estimator $\hat{\bar{Y}}_r$ can be approximated by the difference estimator

$$\tilde{\bar{Y}}_r = \hat{\bar{Y}} + (\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}})' \tilde{\boldsymbol{\beta}}, \tag{2}$$

where $\tilde{\boldsymbol{\beta}}$ is the limit in probability of the vector $\hat{\boldsymbol{\beta}}$, when both the sample size and the population size go to infinity, and the limit is defined as in Isaki and Fuller (1992): Wright (1983); Montanari (1987). Then, the large sample performance of the regression estimator can be studied by means of its linear approximation (2). As a consequence, the regression estimator $\hat{\bar{Y}}_r$ is approximately unbiased, because $\tilde{\bar{Y}}_r$ is unbiased. The sampling variance of $\hat{\bar{Y}}_r$ can be approximated by that of $\tilde{\bar{Y}}_r$ given by

$$V(\tilde{\bar{Y}}_r) = V(\hat{\bar{Y}}) + \tilde{\boldsymbol{\beta}}' V(\hat{\bar{\mathbf{X}}}) \tilde{\boldsymbol{\beta}} - 2\, \tilde{\boldsymbol{\beta}}' C(\hat{\bar{\mathbf{X}}}, \hat{\bar{Y}}), \tag{3}$$

where $V(\hat{\bar{Y}})$ is the variance of $\hat{\bar{Y}}$, $V(\hat{\bar{\mathbf{X}}})$ is the $q \times q$ dimensional variance matrix of $\hat{\bar{\mathbf{X}}}$, and $C(\hat{\bar{\mathbf{X}}}, \hat{\bar{Y}})$ is the q dimensional covariance vector between $\hat{\bar{\mathbf{X}}}$ and $\hat{\bar{Y}}$. Since $\tilde{\bar{Y}}_r$ can be rewritten

$$\tilde{\bar{Y}}_r = \bar{\mathbf{X}}' \tilde{\boldsymbol{\beta}} + \sum_{i\in s} \frac{U_i}{N\pi_i},$$

[1] Giorgio E. Montanari, Dipartimento di Scienze Statistiche, Università di Perugia, Via A. Pascoli - 06100 Perugia, Italy.

where $U_i = Y_i - \boldsymbol{x}_i' \bar{\boldsymbol{\beta}}$, then

$$V(\tilde{\bar{Y}}_r) = \sum_{i=1}^{N} U_i^2 \frac{1 - \pi_i}{N^2 \pi_i} + \sum_{i=1}^{N} \sum_{j \neq 1}^{N} U_i U_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j}.$$

An approximately unbiased estimator of $V(\tilde{\bar{Y}}_r)$ is given by the Horvitz-Thompson formula

$$\hat{V}(\tilde{\bar{Y}}_r) = \sum_{i \in s} \hat{U}_i^2 \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{i \in s} \sum_{j \neq i} \hat{U}_i \hat{U}_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j \ \pi_{ij}},$$

where $\hat{U}_i = Y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$. Alternatively, when the sample size is fixed, the Yates-Grundy variance estimator is available, *i.e.*

$$\hat{V}(\tilde{\bar{Y}}_r) = \sum_{i \in s} \sum_{j > i} \frac{(\pi_i \pi_j - \pi_{ij})}{N^2 \pi_{ij}} \left(\frac{\hat{U}_i}{\pi_i} - \frac{\hat{U}_j}{\pi_j} \right)^2 .$$

Henceforth $V(\tilde{\bar{Y}}_r)$ will be called asymptotic variance of $\hat{\bar{Y}}_r$.

3. THE GENERALIZED REGRESSION ESTIMATOR

Two methods are generally used for constructing the vector $\hat{\boldsymbol{\beta}}$. The first one has been developed within the framework of the model assisted approach to survey sampling inference, as it is described in Särndal, Swensson and Wretman (1992; sec. 6.4) and Estevao, Hidiroglou and Särndal (1995). Letting Y_i be either a random variable or an observation of it, consider the following linear regression superpopulation model

$$\begin{cases} E_m(Y_i) = \boldsymbol{x}_i' \boldsymbol{\beta}, \quad i = 1, 2, \dots, N, \\ V_m(Y_i) = \sigma^2 v_i, \\ C_m(Y_i, Y_j) = 0, \quad i \neq j, \end{cases} \tag{4}$$

where E_m, V_m and C_m denote expected value, variance and covariance with respect to the model; $\boldsymbol{\beta}$ and σ^2 are unknown model parameters; v_i is a known function of $\boldsymbol{x}_i$. The vector

$$\bar{\boldsymbol{\beta}}_1 = \left[\sum_{i=1}^{N} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{v_i} \right]^{-1} \sum_{i=1}^{N} \frac{\boldsymbol{x}_i Y_i}{v_i}$$

is the census least squares estimator of $\boldsymbol{\beta}$. Under general conditions, such as those quoted in the referenced papers,

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_{i \in s} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i v_i} \right]^{-1} \sum_{i \in s} \frac{\boldsymbol{x}_i Y_i}{\pi_i \ v_i}, \tag{5}$$

is a consistent estimator of $\bar{\boldsymbol{\beta}}_1$ and when replaced in (1) gives the generalized regression (GREG) estimator

$$\hat{\bar{Y}}_{r1} = \hat{\bar{Y}} + (\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})' \hat{\boldsymbol{\beta}}_1. \tag{6}$$

In addition to those stated in section 2, this estimator has the following properties: (i) the means of the auxiliary variables estimated through GREG equal the corresponding known population means, *i.e.* $\hat{\bar{\boldsymbol{X}}}_{r1} = \bar{\boldsymbol{X}}$; (ii) the model expected value of the asymptotic sampling variance, *i.e.* $E_m V(\tilde{\bar{Y}}_{r1})$, is a minimum among all asymptotically design-unbiased estimators of $\bar{Y}$ (Wright 1983). Consequently, if the model is well specified, no other asymptotically unbiased estimator exists that is on the average (with respect to the model) more efficient than $\hat{\bar{Y}}_{r1}$.

Well known estimators currently used in practice, such as the ratio and post-stratified estimator, belong to the class of GREG estimators. Furthermore, such a class has recently been extended by means of the calibration technique (Deville and Särndal 1992) to better control the variability of the final observation weights.

4. THE OPTIMAL ESTIMATOR

For constructing an alternative regression estimator based on the same auxiliary variable $\boldsymbol{x}_i$ a second approach considers the vector $\bar{\boldsymbol{\beta}}$ that minimizes the asymptotic variance (3) of the difference estimator (2). Assuming $V(\hat{\bar{\boldsymbol{X}}})$ non singular, *i.e.* there are no linear combinations of the entries of $\hat{\bar{\boldsymbol{X}}}$ with a zero sampling variance, the minimum variance vector is given by

$$\bar{\boldsymbol{\beta}}_2 = [V(\hat{\bar{\boldsymbol{X}}})]^{-1} C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}).$$

Now, consider the unbiased estimators $\hat{V}(\hat{\bar{\boldsymbol{X}}})$ and $\hat{C}(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})$ of $V(\hat{\bar{\boldsymbol{X}}})$ and $C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})$, respectively, that exist provided that the second order inclusion probabilities of the sample design are all positive. They are given by the Horvitz-Thompson formula or the Yates-Grundy formula when applicable. For example, using the former we have the estimated covariance vector

$$\hat{C}(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}) = \sum_{i \in s} \boldsymbol{x}_i Y_i \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{i \in s} \sum_{j \neq i} \boldsymbol{x}_i Y_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j \ \pi_{ij}}.$$

Using $\hat{V}(\hat{\bar{\boldsymbol{X}}})$ and $\hat{C}(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})$ we get the alternative regression estimator

$$\hat{\bar{Y}}_{r2} = \hat{\bar{Y}} + (\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})' \hat{\boldsymbol{\beta}}_2,$$

where $\hat{\boldsymbol{\beta}}_2 = [\hat{V}(\hat{\bar{\boldsymbol{X}}})]^{-1} \hat{C}(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})$. It was studied by Montanari (1987) and called by Rao (1994) the optimal estimator. When $V(\hat{\bar{\boldsymbol{X}}})$ is singular and its rank is $q' < q$, to define the optimal estimator it is understood that one or more entries of $\boldsymbol{x}_i$, hence of $\hat{\bar{\boldsymbol{X}}}$ have to be dropped in such a way as to obtain a $q' \times q'$ non singular variance matrix.

Using the expression for $\tilde{\boldsymbol{\beta}}_2$, the asymptotic variance of $\hat{\bar{Y}}_{r2}$ simplifies to

$$V(\tilde{\bar{Y}}_{r2}) = V(\hat{\bar{Y}}) - C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})'[V(\hat{\bar{\boldsymbol{X}}})]^{-1} C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}). \qquad (7)$$

The properties of the optimal estimator are: (i) asymptotically, the efficiency of $\hat{\bar{Y}}_{r2}$ is not inferior to that of $\hat{\bar{Y}}_{r1}$, *i.e.*, $V(\tilde{\bar{Y}}_{r2}) \leq V(\tilde{\bar{Y}}_{r1})$; (ii) the means of the auxiliary variables estimated through the optimal estimator equal the corresponding known population means, *i.e.* $\hat{\bar{\boldsymbol{X}}}_{r2} = \bar{\boldsymbol{X}}$. As for the case of the GREG estimator, when there is more than one survey variable, the optimal estimator $\hat{\bar{Y}}_{r2}$ can be expressed as a simple weighted estimator with the same weights applying to all variables of interest. For example, using the Horvitz-Thompson formula for variance and covariance estimators, we can write $\hat{\bar{Y}}_{r2} = \sum_{i \in s} Y_i w_i$ where

$$w_i = \frac{1}{\pi_i} + (\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})' [\hat{V}(\hat{\bar{\boldsymbol{X}}})]^{-1} \left(\boldsymbol{x}_i \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{\substack{j \neq i \\ j \in s}} \boldsymbol{x}_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j \pi_{ij}} \right).$$

A similar result can be achieved with the Yates-Grundy formula.

Note that the asymptotic optimality of $\hat{\bar{Y}}_{r2}$ is a strictly design based property, achieved conditionally on the realized finite population (hence, within the fixed population approach to the finite population inference). On the contrary, the asymptotic optimality of $\hat{\bar{Y}}_{r1}$ requires the model to be true, and concerns the average asymptotic variance over the finite populations that can be generated under the model.

Because of these results, $\hat{\bar{Y}}_{r2}$ would seem preferable to $\hat{\bar{Y}}_{r1}$. However, $\hat{\boldsymbol{\beta}}_1$ is a function of population total estimators, and $\hat{\boldsymbol{\beta}}_2$ is a function of variance and covariance estimators. As a consequence, the former is more vulnerable to model misspecification, and the latter is more vulnerable to sampling fluctuations. In a finite size sample, $\hat{\bar{Y}}_{r2}$ is generally less stable and more complex to compute and its variance can be greater than that of $\hat{\bar{Y}}_{r1}$; see Casady and Valliant (1993). However, if an adequate number, g, of degrees of freedom are available for estimating $\boldsymbol{\beta}_2$, the instability problem of $\hat{\bar{Y}}_{r2}$ can be overcome. For example, for standard complex sampling designs having with-replacement sampling at the first stage, g can be roughly taken as the number of sample clusters minus the number of strata (Lehtonen and Pahkinen 1995; p. 181; see Eltinge and Jang 1996, for more elaboration on this topic). A stable $\hat{\boldsymbol{\beta}}_2$ can be expected when g is large enough relative to the dimension q of the auxiliary variable $\boldsymbol{x}_i$. Since with modern computers the computation of $\hat{\bar{Y}}_{r2}$ is less problematic, it becomes interesting to develop a criterion for recognizing when such an estimator is truly advantageous.

5. A CRITERION FOR CHOOSING BETWEEN $\hat{\bar{Y}}_{r1}$ AND $\hat{\bar{Y}}_{r2}$

Consider the following theorem:

Theorem: Let $V(\tilde{\bar{Y}}_r)$ and $V(\tilde{\bar{Y}}_{r2})$ be the asymptotic variances of the general regression estimator $\hat{\bar{Y}}_r$ and the optimal estimator $\hat{\bar{Y}}_{r2}$, respectively. Then

$$V(\tilde{\bar{Y}}_r) - V(\tilde{\bar{Y}}_{r2}) = C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_r)'[V(\hat{\bar{\boldsymbol{X}}})]^{-1} C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_r). \qquad (8)$$

Proof: Using (3) and (7), the difference in variances is

$$V(\tilde{\bar{Y}}_r) - V(\tilde{\bar{Y}}_{r2}) = \tilde{\boldsymbol{\beta}}' V(\hat{\bar{\boldsymbol{X}}}) \tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}' C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}) +$$

$$C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})' [V\hat{\bar{\boldsymbol{X}}}]^{-1} C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}).$$

Since $\tilde{\boldsymbol{\beta}}_2 = [V(\hat{\bar{\boldsymbol{X}}})]^{-1} C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}})$ and $\tilde{\boldsymbol{\beta}}' C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}) = \tilde{\boldsymbol{\beta}}' V(\hat{\bar{\boldsymbol{X}}}) \tilde{\boldsymbol{\beta}}_2$ we have

$$V(\tilde{\bar{Y}}_r) - V(\tilde{\bar{Y}}_{r2}) = (\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_2)' V(\hat{\bar{\boldsymbol{X}}}) (\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_2).$$

But, $C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_r) = C(\hat{\bar{\boldsymbol{X}}}, \hat{\bar{Y}}) - V(\hat{\bar{\boldsymbol{X}}})\tilde{\boldsymbol{\beta}} = V(\hat{\bar{\boldsymbol{X}}})(\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}})$ and (8) follows.

Note that the right hand side of (8) is a positive definite quadratic form and it is equal to zero if and only if $C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_r) = \boldsymbol{0}$. Therefore, the smaller the absolute values of the entries of $C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_r)$ are, the smaller the difference $V(\tilde{\bar{Y}}_r) - V(\tilde{\bar{Y}}_{r2})$ is. The main conclusion the theorem provides us is that an efficient use of any known auxiliary variable population mean requires us to adopt estimators that are uncorrelated with the auxiliary variable mean estimator.

Applying the theorem to the GREG estimator, let us consider the k-th entry of $C(\hat{\bar{\boldsymbol{X}}}, \tilde{\bar{Y}}_{r1})$ that can be written

$$C(\hat{\bar{X}}_k, \tilde{\bar{Y}}_{r1}) = \sum_{i=1}^{N} U_i x_{ki} \frac{1 - \pi_i}{N^2 \pi_i} +$$

$$\sum_{i=1}^{N} \sum_{j \neq j}^{N} U_i x_{kj} \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j},$$

where $U_i = Y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_1$. If the superpopulation model (4) is well specified, it follows that $E_m(U_i) = 0$, for all i, and $E_m[C(\hat{\bar{X}}_k, \tilde{\bar{Y}}_{r1})] = 0$. Therefore, $C(\hat{\bar{X}}_k, \tilde{\bar{Y}}_{r1})$ must be approximately zero for all $k = 1, 2, \ldots, q$, being proportional to a weighted average of N uncorrelated random variables with expected values zero. Consequently the difference $V(\tilde{\bar{Y}}_{r1}) - V(\tilde{\bar{Y}}_{r2})$ must be negligable. The result suggests using the more practical $\hat{\bar{Y}}_{r1}$. The conclusion is that the estimator $\hat{\bar{Y}}_{r2}$ can achieve substantial gains in efficiency compared to $\hat{\bar{Y}}_{r1}$ if the superpopulation model upon which the latter is based is not good enough. This can happen because of the specification of the linear superpopulation model is being confined to regressors with a known population mean.

Since the following quantity

$$\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = C(\hat{\bar{X}}, \tilde{\bar{Y}}_{r1})'[V(\hat{\bar{X}})]^{-1}C(\hat{\bar{X}}, \tilde{\bar{Y}}_{r1})/V(\tilde{\bar{Y}}_{r1})$$

gives the asymptotic relative gain in efficiency that can be achieved with $\hat{\bar{Y}}_{r2}$ compared to $\hat{\bar{Y}}_{r1}$, we propose it as an indicator of a model inadequacy for extracting all information from the sample. When $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ is greater than 10% or 15%, say, the optimal estimator should be adopted. Provided that the second order inclusion probabilities are all positive, under general conditions $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ can be consistently estimated from sample data. Then, the information offered by the estimate $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ can be used for shifting from $\hat{\bar{Y}}_{r1}$ to $\hat{\bar{Y}}_{r2}$ in the next repetition of a periodic survey, or, as we suggest in section 6, within the same survey, choosing between $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ at the estimation stage.

This section concludes with a few examples.

Example 1. Consider a simple random sample of n units and the linear regression model through the origin $E_m(Y_i) = x_i\beta$, $V_m(Y_i) = \sigma^2 x_i$, $C_m(Y_i, Y_j) = 0$, $i \neq j$, assuming $\bar{X}$ known. In this case the GREG is the ratio estimator of the mean, *i.e.*, $\hat{\bar{Y}}_{r1} = \bar{X}\bar{y}/\bar{x}$, where $\bar{y}$ and $\bar{x}$ are the sample means of y and x, respectively. The linear approximation is $\tilde{\bar{Y}}_{r1} = \bar{X}R + \sum_{i\in s} U_i/n$, where $U_i = Y_i - Rx_i$ and $R = \bar{Y}/\bar{X}$. Then, the covariance of $\bar{x}$ and $\tilde{\bar{Y}}_{r1}$ is

$$C(\bar{x}, \tilde{\bar{Y}}_{r1}) = \frac{N-n}{Nn} S_x^2 \left[\frac{S_{xy}}{S_x^2} - R\right], \tag{9}$$

where S_{xy} is the population covariance between y and x and S_x^2 is the population variance of x. If the model is well specified; then $S_{yx}/S_x^2 \approx R$ and expression (9) must be approximately zero. Otherwise, the greater the absolute value of an intercept in a census linear regression of y on x, the more $\hat{\bar{Y}}_{r2}$ is asymptotically efficient than $\hat{\bar{Y}}_{r1}$. The result is not new (for example, see Cochran 1977; sec. 7.5), but it is achieved within the framework of a general theory. Note that $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = [S_{xy}/S_x^2 - R]^2 S_x^2/S_u^2$, where S_u^2 is the population variance of U_i, is a constant with respect to the sample size. When $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ is not negligable, $\hat{\bar{Y}}_{r2}$ should be chosen as regression estimator, or, alternatively, an intercept plugged into the model in order to use the corresponding GREG estimator $\hat{\bar{Y}}_{r1}$. However, for simple random sampling both solutions give the same estimator, *i.e.*, $\hat{\bar{Y}}_{r1} = \hat{\bar{Y}}_{r2}$, but in general they are different, even for self-weighting designs.

Example 2. Consider a stratified random sample and the linear homoscedastic regression model $E_m(Y_i) = \alpha + x_i\beta$, $V_m(Y_i) = \sigma^2$, $C_m(Y_i, Y_j) = 0$, $i \neq j$. Assume that $\bar{X}$ is known and that individual x_i's are known only for sample units and not for the nonsampled units. Now, the auxiliary information is given by $\boldsymbol{x}_i = (1, x_i)'$ and the corresponding GREG estimator can be written $\hat{\bar{Y}}_{r1} = \hat{\bar{Y}} + (\bar{X} - \hat{\bar{X}})\hat{\beta}_1$, where

$$\hat{\beta}_1 = \frac{(\sum_{i\in s} Y_i x_i/N\pi_i) - \hat{\bar{X}}\hat{\bar{Y}}}{(\sum_{i\in s} x_i^2/N\pi_i) - \hat{\bar{X}}^2},$$

and where the estimated α cancels out. Because $\tilde{\beta}_1 = S_{yx}/S_x^2$ and $U_i = Y_i - \bar{Y} - \tilde{\beta}_1(x_i - \bar{X})$, we have

$$C(\hat{\bar{X}}, \tilde{\bar{Y}}_{r1}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{N^2 n_h} S_{hx}^2(\tilde{\beta}_{h1} - \tilde{\beta}_1), \tag{10}$$

where the subindex h denotes stratum quantities and $\tilde{\beta}_{h1} = S_{hxy}/S_{hx}^2$. The right hand side of (10) is a function of the differences between each within-stratum regression coefficient and the coefficient for the whole population. If the model is well specified, the differences $\tilde{\beta}_{h1} - \tilde{\beta}_1$ must be negligible. Otherwise, $C(\hat{\bar{X}}, \tilde{\bar{Y}}_{r1})$ can take non negligible absolute values and, since only $\bar{X}$ is known, the estimator $\hat{\bar{Y}}_{r2}$ appears to extract better all the information from the sample value of $\hat{\bar{X}}$

It is interesting to note that when the allocation of the sample is proportional, *i.e.*, $n_h \propto N_h$, ignoring terms of order $1/N_h$ relative to unity, $\hat{\bar{Y}}_{r2}$ is equal to the GREG estimator based on the auxiliary variable $\boldsymbol{x}_i = (d_{1i}, d_{2i}, \ldots, d_{Hi}, x_i)'$ and $v_i = 1$, where d_{hi} is an indicator variable of the membership of unit i to stratum $h = 1, 2, \ldots, H$. This model fits different regression lines with a common slope within the strata.

Example 3. Consider a complex sampling design and suppose that the population can be partitioned into H post-strata of known sizes. Assume the superpopulation model $E_m(Y_i) = \beta_{h(i)}$, $V_m(Y_i) = \sigma^2$, and $C_m(Y_i, Y_j) = 0$, $i \neq j$, where the subindex $h(i)$ denotes the post-stratum to which the i-th unit belongs. Denoting by d_{hi} the indicator variable of the i-th unit membership to post-stratum h, and with $\bar{D}_h$ its known population mean, putting $\boldsymbol{x}_i = (d_{1i}, d_{2i}, \ldots, d_{Hi})'$ and $v_i = 1$, in (5), we get the post-stratified estimator, $\hat{\bar{Y}}_{r1} = \sum_1^H \bar{D}_h \hat{\bar{Z}}_h/\hat{\bar{D}}_h$, where $\hat{\bar{Z}}_h$ and $\hat{\bar{D}}_h$ are the Horvitz-Thompson mean estimators of the variables $z_{hi} = Y_i d_{hi}$ and d_{hi}, respectively. The linear approximation is $\tilde{\bar{Y}}_r = \hat{\bar{Y}} + (\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}})'\boldsymbol{\beta}_1$, where $\tilde{\boldsymbol{\beta}}_1 = (R_1, R_2, \ldots, R_H)'$, $R_h = \bar{Z}_h/\bar{D}_h$ (*i.e.*, the mean value of y in the h-th post-stratum), and $\hat{\bar{\boldsymbol{X}}} = (\hat{\bar{D}}_1, \hat{\bar{D}}_2, \ldots, \hat{\bar{D}}_H)'$. Since $U_i = Y_i - \sum_1^H R_h d_{hi}$, the covariance of $\hat{\bar{D}}_h$ and $\tilde{\bar{Y}}_{r1}$ is

$$C(\tilde{\bar{Y}}_{r1}, \hat{\bar{D}}_h) = C(\hat{\bar{Y}}, \hat{\bar{D}}_h) - \sum_{j=1}^{H} R_j C(\hat{\bar{D}}_j, \hat{\bar{D}}_h). \tag{11}$$

Under the superpopulation model upon which $\hat{\bar{Y}}_{r1}$ is based on, we have $E_m[C(\tilde{\bar{Y}}_{r1}, \hat{\bar{D}}_h)] = 0$ and a negligible value of $C(\tilde{\bar{Y}}_{r1}, \hat{\bar{D}}_h)$ is expected for all h. It can be easily seen that for simple random sampling, formula (11) is identically zero. But in complex sampling schemes such covariances might take non negligible values, for example, when in a multistage sampling scheme a linear regression of the primary unit totals of $z_{hi}^{\cdot}$ on the totals of d_{hi} yields a non negligible intercept for some h. See Casady and Valliant (1993) for a case study.

6. EMPIRICAL STUDIES

The above analysis is based on first order approximations. In the following empirical studies the finite sample performances of $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ will be explored within the framework of example 2.

6.1 The First Empirical Study

In this first empirical study we consider a population of infinite size subdivided into two strata of equal weights and a proportional stratified random sampling design to estimate the mean of a survey variable y. To this end, let us suppose that there exists a scalar variable x that was not available for stratification but with a known population mean $\bar{X}$ and unknown stratum means (*i.e.*, the x values are not available for nonsampled units).

Since only the population mean of x is assumed known, a reasonable superpopulation model that can be assumed to identify a GREG estimator is the linear regression one, with homoscedastic errors, *i.e.*, $E_m(Y_i) = \alpha + x_i\beta$, $V_m(Y_i) = \sigma^2$, $C_m(Y_i, Y_j) = 0, i \neq j$. The auxiliary variable plugged into (5) is $\boldsymbol{x}_i = (1, x_i)'$ and the corresponding GREG estimator can be written

$$\hat{\bar{Y}}_{r1} = \bar{y} + (\bar{X} - \bar{x}) s_{yx} / s_x^2,$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y and x, s_{yx} is the sample covariance between y and x, and s_x^2 is the sample variance of x. The linear approximation is

$$\tilde{\bar{Y}}_{r1} = \bar{y} + (\bar{X} - \bar{x}) S_{yx} / S_x^2,$$

where S_{yx} and S_x^2 are the population analogues of s_{yx} and s_x^2.

Dropping the first component of $\boldsymbol{x}_i = (1, x_i)'$, whose mean is estimated without error, the optimal estimator based on the same auxiliary variable is given by

$$\hat{\bar{Y}}_{r2} = \bar{y} + (\bar{X} - \bar{x}) \hat{C}(\bar{y}, \bar{x}) / \hat{V}(\bar{x}),$$

where $\bar{X}$ is the population mean of x, $\hat{C}(\bar{y}, \bar{x})$ and $\hat{V}(\bar{x})$ are the standard unbiased estimators of the covariance between $\bar{y}$ and $\bar{x}$ and the variance of $\bar{x}$, respectively. The corresponding linear approximation is

$$\tilde{\bar{Y}}_{r2} = \bar{y} + (\bar{X} - \bar{x}) C(\bar{y}, \bar{x}) / V(\bar{x}),$$

where $C(\bar{y}, \bar{x})$ and $V(\bar{x})$ are the true covariance and variance.

In this case, the expression of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ simplifies to

$$\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = \frac{\sum_1^2 S_{hx}^2}{\sum_1^2 S_{hu}^2} \left(\frac{\sum_1^2 S_{hxy}}{\sum_1^2 S_{hx}^2} - \frac{S_{yx}}{S_x^2} \right),$$

and it can be estimated replacing the population variances and covariances with the sample analogues.

Four simulations were performed. In the first two, the sample values of x were drawn from a uniform distribution on [30-70] in the first stratum and [50-90] in the second one. The sample values of y, given x, were drawn from a normal distribution with expected values $1.26x$ in the first stratum and $0.82x$ in the second. The conditional variance was $8x$ in both strata in the first simulation and $3x$ in the second one. In the third and fourth simulation, the sample values of x were drawn from a linearly transformed gamma random variable with parameters chosen to achieve the first two simulation stratum means and variances for x and y and an asymmetry index for x (given by the ratio between the third central moment and the third power of the standard deviation) equal to 2.5. This allows studying the effects of a strong asymmetry in the marginal distributions of y and x.

The populations were constructed to have $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) =$ 8.1% when $V(Y|x) = 8x$, and $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = 18.6\%$, when $V(Y|x) = 3x$. Note that the GREG estimator based on the true model is the separate ratio estimator; however, its use would require the knowledge of the stratum means of x, but they are assumed unknown.

In each simulation we drew 10,000 samples of size 20 (ten units per stratum), and 5,000 of size 40 (twenty units per stratum). For each sample we computed the values of the Horvitz-Thompson estimator $\hat{\bar{Y}} = \bar{y}$, and of $\tilde{\bar{Y}}_{r1}, \tilde{\bar{Y}}_{r2}$, $\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}$, and $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. We also computed an estimator $\hat{\bar{Y}}_{r3}$, defined to take the value of $\hat{\bar{Y}}_{r1}$, when $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) \leq$ 8%, and the value of $\hat{\bar{Y}}_{r2}$ otherwise. So, $\hat{\bar{Y}}_{r3}$ is a sample dependent type estimator, constructed choosing between $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ according to the estimated value of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. Here, 8% is an arbitrarily chosen threshold, over which shifting from $\hat{\bar{Y}}_{r1}$ to $\hat{\bar{Y}}_{r2}$ is thought to be convenient.

Table 1 reports for each simulation the empirical results achieved with reference to the percent relative bias of estimators (RB) and the mean squared error (MSE), in the latter case having set that of the Horvitz-Thompson estimators equal to 100 by multiplying the MSE values by $100/\text{MSE}(\bar{y})$. As we can see, the biases are all negligible (the biggest absolute value is less than 0.6% and all biases are less than 10% of the corresponding standard errors) and contribute to the MSE in a negligible manner. The MSE reduction percentages that can be achieved shifting from $\tilde{\bar{Y}}_{r1}$ to $\tilde{\bar{Y}}_{r2}$ are approximately equal to the fixed in advance values of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$, *i.e.*, 8.1% and 18.6%. The effective MSE values of $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ are greater than the corresponding asymptotic values, in particular when the population is asymmetric and the estimator is the optimal one. For example, in the third simulation, when $n = 20$, the MSE of $\hat{\bar{Y}}_{r1}$ shows a 5.1% relative increase compared to that of $\tilde{\bar{Y}}_{r1}$, while the corresponding value for $\hat{\bar{Y}}_{r2}$ is 10.7%. Doubling the sample size, those relative values decrease to 2.8% and 3.6%, respectively. As we observed in example 2, when the sample allocation is proportional, $\hat{\bar{Y}}_{r2}$ is equal to the GREG estimator based on a homoscedastic linear model that fits two parallel regression lines in the two strata. So, the greater loss in efficiency percentage of $\hat{\bar{Y}}_{r2}$ with respect to its asymptotic variance can be explained by the added parameter to be estimated in the model.

The performance of $\hat{\bar{Y}}_{r3}$ is also interesting; this estimator is approximately unbiased and its MSE is lower than that of $\hat{\bar{Y}}_{r1}$ the more often $\hat{\bar{Y}}_{r2}$ is selected. Table 1 reports for each simulation the percentages of samples for which $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) > 8\%$ and $\hat{\bar{Y}}_{r2}$ was selected instead of $\hat{\bar{Y}}_{r1}$. The higher is the theoric value of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$, the more often $\hat{\bar{Y}}_{r2}$ is chosen over $\hat{\bar{Y}}_{r1}$.

Obviously, the performance of $\hat{\bar{Y}}_{r3}$ depends on the sampling distribution of the sample statistics $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. Table 2 reports the means, the standard deviations, and some quantiles of the empirical distributions of $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ for the gamma populations, which are the more problematic ones. As it can be seen, the distributions of $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ were in all cases positively skewed and highly variable. This means that larger sample sizes than those considered here are needed to get reliable estimates of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. Clearly, the less the variance of $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$, the higher is the gain in efficiency of $\hat{\bar{Y}}_{r3}$ over $\hat{\bar{Y}}_{r1}$ when the true value of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ is over the threshold for $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ chosen to shift from $\hat{\bar{Y}}_{r1}$ to $\hat{\bar{Y}}_{r2}$.

Table 1

Empirical percent relative bias (RB) and Mean Squared Error (MSE) of $\bar{y}$, $\tilde{\bar{Y}}_{r1}$, $\tilde{\bar{Y}}_{r2}$, $\hat{\bar{Y}}_{r1}$, $\hat{\bar{Y}}_{r2}$ and $\hat{\bar{Y}}_{r3}$ and percentage of samples for which $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) > 8\%$ in the first empirical study

	Uniform populations							
	$V(Y\|x) = 8x$				$V(Y\|x) = 3x$			
	$n = 20$		$n = 40$		$n = 20$		$n = 40$	
Estimator	RB (%)	MSE	RB (%)	MSE	RB (%)	MSE	RB (%)	MSE
$\bar{y}$	-0.06	100.0	-0.08	100.0	0.12	100.0	-0.10	100.0
$\tilde{\bar{Y}}_{r1}$	-0.05	83.8	-0.06	84.1	0.10	69.4	-0.05	68.8
$\tilde{\bar{Y}}_{r2}$	-0.03	77.3	-0.04	77.7	0.07	56.2	0.01	55.8
$\hat{\bar{Y}}_{r1}$	0.07	87.7	-0.01	86.2	0.22	73.4	-0.00	70.5
$\hat{\bar{Y}}_{r2}$	-0.05	82.4	-0.04	80.1	0.05	59.8	-0.00	57.3
$\hat{\bar{Y}}_{r3}$	-0.06	85.0	-0.05	83.1	0.03	61.0	-0.01	57.9
Freq ($\lambda > 8\%$)	53.5%		53.6%		88.6%		93.5%	
	Gamma populations							
	$V(Y\|x) = 8x$				$V(Y\|x) = 3x$			
	$n = 20$		$n = 40$		$n = 20$		$n = 40$	
Estimator	RB(%)	MSE	RB(%)	MSE	RB(%)	MSE	RB(%)	MSE
$\bar{y}$	0.07	100.0	-0.01	100.0	0.02	100.0	-0.03	100.0
$\tilde{\bar{Y}}_{r1}$	0.08	84.1	0.02	84.3	0.06	69.8	-0.03	69.9
$\tilde{\bar{Y}}_{r2}$	0.09	77.5	0.05	78.1	0.10	57.1	-0.02	56.9
$\hat{\bar{Y}}_{r1}$	-0.58	88.4	-0.30	86.7	-0.60	75.5	-0.36	72.8
$\hat{\bar{Y}}_{r2}$	0.03	85.8	0.03	80.9	0.12	63.5	-0.02	59.1
$\hat{\bar{Y}}_{r3}$	-0.05	87.9	0.07	86.2	0.06	65.4	-0.04	60.8
Freq ($\lambda > 8\%$)	50.6%		50.3%		86.9%		91.7%	

Table 2

Selected characteristics of the empirical distributions of $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ for gamma populations (first empirical study)

Gamma Populations	Mean	Standard deviation	Median	Quantiles 10%	Quantiles 90%
$V(Y\|x) = 8x,\ n = 20$	10.7	9.8	8.7	1.3	24.9
$V(Y\|x) = 8x,\ n = 40$	9.2	6.3	8.3	2.5	19.1
$V(Y\|x) = 3x,\ n = 20$	21.6	12.3	19.2	6.9	40.7
$V(Y\|x) = 3x,\ n = 40$	19.0	9.5	18.9	9.4	34.2

6.2 The Second Empirical Study

In the second empirical study, we consider a finite population subdivided into eight strata each of size 100, according to an auxiliary variable x whose values are assumed known for each unit of the population. In order to simulate a stratification based on x, the values of x were assigned through the monotonic function of h and i

$$x_{hi} = 4.95 + 5\sum_{j=1}^{h-1} j + h \cdot i,$$

where hi is the label of the unit $i = 1, 2, \ldots, 100$ within the stratum $h = 1, 2, \ldots, 8$.

. A finite population of y values, given x, was generated using the model

$$Y_{hi} = 20 + 2x_{hi} + 0.06x_{hi}^2 + \epsilon_{hi} \cdot x_{hi},$$

where ϵ_{hi} is a standard normal random variable. The realized values of the mean, standard deviation and asymmetry index of y are 618.2, 676.0, and 1.21, respectively. The correlation between y and x is 0.96.

A proportional stratified random sampling without replacement design was used to select 5,000 samples of size $n = 40$ (five units per stratum) and 2,500 samples of size 80 (ten units per stratum). For each sample we computed the following quantities:

- the unbiased estimator of the population mean $\bar{Y}$, *i.e.*, $\bar{y}$;
- the ratio estimator $\hat{\bar{Y}}_{r11}$, based on the model $E_m(Y_{hi}) = \beta x_{hi}$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}$, and obtained from (5) and (6) putting $\boldsymbol{x}_{hi} = x_{hi}$ and $v_{hi} = x_{hi}$;
- the optimal estimator $\hat{\bar{Y}}_{r21}$, based on the same auxiliary variable used for $\hat{\bar{Y}}_{r11}$;
- the GREG estimator $\hat{\bar{Y}}_{r12}$, based on the model $E_m(Y_{hi}) = \alpha + \beta x_{hi}$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}$, and obtained from (5) and (6) putting $\boldsymbol{x}_{hi} = (1, x_{hi})'$ and $v_{hi} = x_{hi}$;
- the optimal estimator $\hat{\bar{Y}}_{r22}$ based on the same auxiliary variables used for $\hat{\bar{Y}}_{r12}$;
- the GREG estimator $\hat{\bar{Y}}_{r13}$, based on the model $E_m(Y_{hi}) = \alpha + \beta x_{hi} + \gamma x_{hi}^2$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}^2$ (the true model), and obtained from (5) and (6) putting $\boldsymbol{x}_{hi} = (1, x_{hi}, x_{hi}^2)'$ and $v_{hi} = x_{hi}^2$;
- the optimal estimator $\hat{\bar{Y}}_{r23}$ based on the same auxiliary variables used for $\hat{\bar{Y}}_{r13}$;
- the linear approximations $\tilde{\bar{Y}}_{r12}$, $\tilde{\bar{Y}}_{r13}$, $\tilde{\bar{Y}}_{r22}$, and $\tilde{\bar{Y}}_{r23}$ of $\hat{\bar{Y}}_{r12}$, $\hat{\bar{Y}}_{r13}$, $\hat{\bar{Y}}_{r22}$, and $\hat{\bar{Y}}_{r23}$, respectively;
- the statistics $\hat{\lambda}(\hat{\bar{Y}}_{r1k}, \hat{\bar{Y}}_{r2k})$, for $k = 1, 2, 3$;
- the sample dependent estimators $\hat{\bar{Y}}_{r3k}$ $(k = 1, 2, 3)$ defined to take the value of $\hat{\bar{Y}}_{r1k}$ when $\hat{\lambda}(\hat{\bar{Y}}_{r1k}, \hat{\bar{Y}}_{r2k}) \leq 8\%$, and the value of $\hat{\bar{Y}}_{r2k}$ otherwise.

We do not consider separate regression estimation because sample sizes within strata are small. The finite population is such that $\lambda(\hat{\bar{Y}}_{r11}, \hat{\bar{Y}}_{r21}) = 0.22$, $\lambda(\hat{\bar{Y}}_{r12}, \hat{\bar{Y}}_{r22}) = 0.16$, and $\lambda(\hat{\bar{Y}}_{r13}, \hat{\bar{Y}}_{r23}) = 0.00$. Note that because of the sample design considered we have $\hat{\bar{Y}}_{r21} = \hat{\bar{Y}}_{r22}$ and therefore we omit $\hat{\bar{Y}}_{r21}$.

Table 3 reports the empirical results achieved with reference to the percent relative bias of estimators (RB) and the Mean Squared Error (MSE), in the latter case having set that of the Horvitz-Thompson estimators equal to 100. The results are separated according to the sample size.

Again, the biases are all negligible. The MSE reduction percentage that can be achieved with respect to the sample mean increases with the number of auxiliary variables used. However, as expected $\hat{\bar{Y}}_{r11}$ and $\hat{\bar{Y}}_{r12}$ are less efficient than the optimal estimator $\hat{\bar{Y}}_{r22}$ based on the same auxiliary variables. The statistics $\hat{\lambda}(\hat{\bar{Y}}_{r11}, \hat{\bar{Y}}_{r21})$ and $\hat{\lambda}(\hat{\bar{Y}}_{r12}, \hat{\bar{Y}}_{r22})$ take values above the 8% threshold most of the time, especially when the sample size is 80. The sample dependent estimators $\hat{\bar{Y}}_{r31}$ and $\hat{\bar{Y}}_{r32}$ are both more efficient than $\hat{\bar{Y}}_{r11}$ and $\hat{\bar{Y}}_{r12}$. The result is due to the inadequacy of the models upon which $\hat{\bar{Y}}_{r11}$ and $\hat{\bar{Y}}_{r12}$ are based for extracting all information from the sample. On the other hand, $\hat{\bar{Y}}_{r13}$ is more efficient than $\hat{\bar{Y}}_{r23}$ because it is based on the true model. Most of the time the statistic $\hat{\lambda}(\hat{\bar{Y}}_{r13}, \hat{\bar{Y}}_{r23})$ is below the threshold, especially when the sample size is 80, and the sample dependent estimator $\hat{\bar{Y}}_{r33}$ is almost as efficient as $\hat{\bar{Y}}_{r13}$.

Looking at the linear approximations, first we observe that the MSE's of the GREG estimators $\hat{\bar{Y}}_{r12}$ and $\hat{\bar{Y}}_{r13}$ are almost equal to those of $\tilde{\bar{Y}}_{r12}$ and $\tilde{\bar{Y}}_{r13}$ in this second study. This is not true for the optimal estimators $\hat{\bar{Y}}_{r22}$ and $\hat{\bar{Y}}_{r23}$. The losses in efficiency with respect to their linear approximations $\tilde{\bar{Y}}_{r22}$ and $\tilde{\bar{Y}}_{r23}$ are greater, but they diminish rapidly when the sample size increases. The MSE's of the linear approximations confirm that given a certain amount of auxiliary information, a negligible gain in efficiency can be achieved through the optimal estimator, even with very large samples (compare $\tilde{\bar{Y}}_{r13}$ with $\tilde{\bar{Y}}_{r23}$), when the model upon which the GREG is based holds true. Substantial gains in efficiency can be achieved if the model is not adequate, such as those upon which $\hat{\bar{Y}}_{r11}$ and $\hat{\bar{Y}}_{r12}$ are based (compare $\tilde{\bar{Y}}_{r12}$ with $\tilde{\bar{Y}}_{r22}$). Table 4 reports the means, standard deviations and some quantiles of the empirical distributions of $\hat{\lambda}(\hat{\bar{Y}}_{r1k}, \hat{\bar{Y}}_{r2k})$, $k = 1, 2, 3$.

Table 3
Empirical percent relative bias (RB) and Mean Squared Error (MSE) of estimators and percentage of samples for which $\hat{\lambda}(\hat{\bar{Y}}_{r1k}, \hat{\bar{Y}}_{r2k}) > 8\%$ in the second empirical study

Auxiliary used	Estimator	Sample size 40			Sample size 80		
		RB(%)	MSE	$(\lambda > 8\%)$	RB(%)	MSE	$(\lambda > 8\%)$
none	$\bar{y}$	0.01	100.0	–	0.01	100.0	–
(x)	$\hat{\bar{Y}}_{r11}$	-0.01	55.2	82.6%	0.00	54.3	85.0%
(x)	$\hat{\bar{Y}}_{r31}$	-0.05	48.4	–	-0.02	43.8	–
$(1, x)'$	$\hat{\bar{Y}}_{r12}$	-0.01	51.7	72.7%	0.00	50.8	83.2%
$(1, x)'$	$\hat{\bar{Y}}_{r22}$	-0.05	47.4	–	-0.01	43.3	–
$(1, x)'$	$\hat{\bar{Y}}_{r32}$	-0.05	48.3	–	-0.02	43.8	–
$(1, x)'$	$\tilde{\bar{Y}}_{r12}$	0.02	51.6	–	0.01	50.7	–
$(1, x)'$	$\tilde{\bar{Y}}_{r22}$	0.02	44.3	–	0.00	42.3	–
$(1, x, x^2)'$	$\hat{\bar{Y}}_{r13}$	-0.01	35.1	28.9%	0.02	33.5	10.5%
$(1, x, x^2)'$	$\hat{\bar{Y}}_{r23}$	-0.10	38.0	–	-0.03	34.7	–
$(1, x, x^2)'$	$\hat{\bar{Y}}_{r33}$	-0.04	37.0	–	-0.01	33.8	–
$(1, x, x^2)'$	$\tilde{\bar{Y}}_{r13}$	0.01	34.9	–	0.03	33.5	–
$(1, x, x^2)'$	$\tilde{\bar{Y}}_{r23}$	0.01	34.7	–	0.03	33.2	–

Table 4
Selected characteristics of the empirical distributions of $\hat{\lambda}(\hat{\bar{Y}}_{r1k}, \hat{\bar{Y}}_{r2k})$, $k = 1, 2, 3$ (second empirical study)

Statistics	Sample size 40					Sample size 80				
	Mean	Standard deviation	Median	Quantiles		Mean	Standard deviation	Median	Quantiles	
				10%	90%				10%	90%
$\hat{\lambda}(\hat{\bar{Y}}_{r11}, \hat{\bar{Y}}_{r21})$	0.24	0.15	0.23	0.04	0.45	0.23	0.10	0.23	0.07	0.35
$\hat{\lambda}(\hat{\bar{Y}}_{r12}, \hat{\bar{Y}}_{r22})$	0.19	0.14	0.17	0.02	0.38	0.18	0.09	0.17	0.04	0.30
$\hat{\lambda}(\hat{\bar{Y}}_{r13}, \hat{\bar{Y}}_{r23})$	0.06	0.08	0.03	0.00	0.18	0.03	0.04	0.01	0.00	0.08

7. DISCUSSION

The optimal estimator can be an efficient alternative to the generalized regression estimator based on misspecified superpopulation models when the sample size is large enough. This efficiency can be measured by means of the sample statistic, $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$, that captures the asymptotic relative gain in efficiency of $\hat{\bar{Y}}_{r2}$ over $\hat{\bar{Y}}_{r1}$, given a certain amount of auxiliary information. The performance of the optimal estimator appears to be good, even in finite size samples, and its use profitable, provided that the value of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ is big enough to compensate for its greater instability. In fact, the empirical results confirm a greater instability in the optimal estimator, especially with asymmetric populations. Further empirical evidence is needed to evaluate its stability when the auxiliary variable is multivariate and to establish when a sample is large enough to overcome the problem.

In order to use the information provided by $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ within the same survey, the distributional properties of this sample statistic and of the sample dependent regression estimator, which seems to perform well in the empirical study, have to be studied in more detail. In particular, the distribution of $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ when its true value is zero will be useful for choosing the threshold over which shifting from $\hat{\bar{Y}}_{r1}$ to $\hat{\bar{Y}}_{r2}$ is truly profitable. Besides working with larger sample sizes, the instability problem of this statistic can be addressed by looking for more stable, consistent estimators of the variances and covariances appearing in $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. Furthermore, since in most practical situations there is more than one variable of interest, in order to apply the same weights to all variable, the optimal estimator should be chosen on the grounds of an averaged λ-measure across the main survey variables, and such an average is more stable than single λ-measures.

ACKNOWLEDGEMENTS

This research was partially funded by a grant of M.U.R.S.T. of Italy. The author is grateful to an associate editor and the referees for their constructive comments which greatly improved upon an earlier version of this paper.

REFERENCES

CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

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

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ELTINGE, J.L., and JANG, D.S. (1996). Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology*, 22, 157-165.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

LEHTONEN, R., and PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Comples Surveys*. New York: Wiley.

MONTANARI, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information a the estimation stage. *Journal of Official Statistics*, 10, 153-165.

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

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 79-88
Statistics Canada

Combining Multiple Frames to Estimate Population Size and Totals

DAWN E. HAINES and KENNETH H. POLLOCK[1]

ABSTRACT

Efficient estimates of population size and totals based on information from multiple list frames and an independent area frame are considered. This work is an extension of the methodology proposed by Hartley (1962) which considers two general frames. A main disadvantage of list frames is that they are typically incomplete. In this paper, we propose several methods to address frame deficiencies. A joint list-area sampling design incorporates multiple frames and achieves full coverage of the target population. For each combination of frames, we present the appropriate notation, likelihood function, and parameter estimators. Results from a simulation study that compares the various properties of the proposed estimators are also presented.

KEY WORDS: Incomplete frame; Capture-recapture sampling; Screening estimator; Dual frame methodology; Multiple frame estimation.

1. INTRODUCTION

In classical sampling theory, it is assumed that a complete frame exists. In practice, however, this assumption is often violated. Frame imperfections such as omissions, duplications, and inaccurate recordings are almost inevitable in any large data collection operation (Hansen, Hurwitz and Madow 1953). Information collected from list and area frames is used to obtain estimates of the unknown population size and totals. For example, an ecologist or wildlife biologist may use one list and one area frame sample to estimate the number of bald eagle nests in a given region. The U.S. Bureau of the Census uses dual system estimation to measure decennial census undercounts. Darroch, Fienberg, Glonek and Junker (1993) describe a three-sample multiple-capture approach to estimating population size when inclusion probabilities are heterogeneous. In addition, state agriculture officials may be interested in estimating the number of hog farms and the total number of hogs in North Carolina. Typically, information from multiple information sources is combined to estimate population sizes and totals.

List frames are physical listings of sampling units in the target population. These are constructed over the years using information from scientists as well as city, county, state, and federal agencies. Items found on a list frame can include, but are not limited to, names, addresses, telephone numbers, social security numbers, or physical descriptions of location. These and other miscellaneous stratification variables are used to identify persons, animals, businesses, or other establishments. When estimating the number of bald eagle nests in a region, we construct this year's list frame using information from last year's list frame. With the addition of new eagle nests, last year's list frame becomes quickly outdated and incomplete. Because of this incompleteness, estimates based solely on list frames typically underestimate the true population size. Supplementing available information with an area frame sample may provide an efficient estimation of the population size and totals.

An area frame is a collection of geographical areas defined by identifiable boundaries. The entire area in which data are collected is divided into mutually exclusive and exhaustive sampling units called segments. The segments are usually stratified according to a characteristic of interest. Once a stratified random sample of segments is drawn, enumerators visit the sampled segments and record measurements on all reporting units contained therein.

The National Agricultural Statistics Service (NASS) currently employs a multi-frame approach for its sampling and estimation of numerous agricultural commodities. 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. Kott and Vogel (1995) provide a general overview of multiple frame surveys.

In Section 2, we consider estimation based on information from two or more independent list frames. We show how these methods are related to capture-recapture methods. In Section 3, we consider more efficient estimators of population size and totals when information from an independent area frame sample is available. We extend these methods to the case of dependent list frames in Section 4. Results from a simulation study that compare different estimators are summarized in Section 5. Finally,

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

Section 6 summarizes our results and discusses future directions for research.

2. MULTIPLE LIST FRAMES

2.1 Population Size Estimation

List frames used to estimate population size are usually incomplete and do not cover the entire population. One solution to the incomplete list frame problem is to merge two or more incomplete list frames. Combining multiple list frames may result in improved coverage of the target population, and thus, may provide better estimators. In the case of multiple list frames, it is commonly assumed that each element in the population has the same probability of being included on a given list frame. Hence, the list frame elements themselves constitute our "samples." For example, individuals may decide independently whether or not to list their telephone numbers in the telephone directory with equal probability. In the case of bald eagle nests, this year's list frame is constructed based on last year's nest sightings. If we assume that the probability of a nest being sighted is the same for all nests, then the above assumption is valid. Finally, the assumption is also valid in capture-recapture experiments where the first list frame consists of all animals captured on the first sampling occasion and the second list frame consists of all animals captured on the second sampling occasion. This scenario corresponds to Model M_t in the capture-recapture literature. See Otis, Burnham, White and Anderson (1978) for details. Model M_t assumes all animals in the population are equally at risk to capture on each sampling occasion, but this probability can vary over different sampling occasions.

To begin, we consider the case of two independent list frames, B_1 and B_2. Suppose B_1 has size N_{B_1} and B_2 has size N_{B_2}. Let domain $b_1 (b_2)$ consist of those $N_{b_1} (N_{b_2})$ elements that belong only to frame $B_1 (B_2)$ and domain $b_1 b_2$ contain $N_{b_1 b_2}$ units that belong to both frames. The final domain includes existing target population elements that are not included on either list frame. Its size is $N - N_{b_1} - N_{b_2} - N_{b_1 b_2}$. Domain notation for list frames B_1 and B_2 is presented in Table 1. Note that every element in every frame must be categorized into a domain without error. Errors in domain determination are serious and cannot be corrected at a later time. These errors are not considered in the estimation phase and thus are regarded as nonsampling errors. Nealon (1984) claims that domain determination is the single largest source of nonsampling error in multiple frame designs (Kott and Vogel 1995).

Let the probability that a population element is included on frame $B_1 (B_2)$ be $p_{B_1} (p_{B_2})$. Since list frames B_1 and B_2 are assumed to be independent, the probability of an element belonging to domain b_1 is $p_{b_1} = p_{B_1}(1 - p_{B_2})$. The remaining domain probabilities are defined similarly. The population size N and the inclusion probabilities p_{B_1} and p_{B_2} are unknown parameters. The likelihood function is given by

$$\mathcal{L}(p_{B_1}, p_{B_2}, N \mid N_{b_1}, N_{b_2}, N_{b_1 b_2}) = \binom{N}{N_{b_1}, N_{b_2}, N_{b_1 b_2}} * p_{B_1}^{N_{B_1}} p_{B_2}^{N_{B_2}} (1 - p_{B_1})^{N - N_{B_1}} (1 - p_{B_2})^{N - N_{B_2}}. \quad (1)$$

Table 1
Domain Notation for List Frames B_1 and B_2

Domain Size	Domain Probability
N_{b_1}	$p_{b_1} = p_{B_1}(1 - p_{B_2})$
N_{b_2}	$p_{b_2} = (1 - p_{B_1})p_{B_2}$
$N_{b_1 b_2}$	$p_{b_1 b_2} = p_{B_1} p_{B_2}$
$N - N_{b_1} - N_{b_2} - N_{b_1 b_2}$	$1 - p_{b_1} - p_{b_2} - p_{b_1 b_2} = (1 - p_{B_1})(1 - p_{B_2})$

Maximum likelihood estimators (MLEs) of the frame inclusion probabilities are obtained by maximizing the logarithm of the likelihood (1). This procedure yields

$$\hat{p}_{B_1} = \frac{N_{B_1}}{\hat{N}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{B_2}}{\hat{N}}, \quad (2)$$

where the MLE $\hat{N}$ is substituted for N. Rather than differentiating the log-likelihood function to approximate the value of N, we employ the "ratio method" of maximizing the likelihood which equates $\mathcal{L}(N)$ to $\mathcal{L}(N - 1)$ (Darroch 1958). This process accounts for the discrete parameter N and yields the equation

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - N_{b_2} - N_{b_1 b_2})} * (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) = 1. \quad (3)$$

Here we assume that N is large so that

$$\frac{N_{B_1}}{N - 1} \approx \frac{N_{B_1}}{N} \quad \text{and} \quad \frac{N_{B_2}}{N - 1} \approx \frac{N_{B_2}}{N}.$$

Substituting the estimators in (2) into (3) yields

$$\hat{N}_1 = \hat{N} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}}. \quad (4)$$

Sekar and Deming (1949) derive an estimate of the variance of (4), given by

$$\hat{V}(\hat{N}_1) = \frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}.$$

Substituting (4) into (2) yields the MLEs of p_{B_1} and p_{B_2},

$$\hat{p}_{B_1} = \frac{N_{b_1 b_2}}{N_{B_2}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{b_1 b_2}}{N_{B_1}}.$$

The estimator $\hat{N}_1$ of N in (4) is called the Lincoln-Petersen estimator in closed population capture-recapture models. The elements on list frame B_1 may be considered as the units captured in the first sampling occasion and the elements on list frame B_2 may be viewed as the units captured in the second sampling occasion. The elements in domain $b_1 b_2$ correspond to recaptured elements. With this correspondence, it is easy to see that the likelihood for the population size and capture probabilities for two occasions will be the same as that given in (1). Hence, the MLEs derived for two independent list frames will be the same as the corresponding MLEs for the capture-recapture model with two sampling occasions.

Extending these ideas, we contend that combining k independent list frames is directly related to having k sampling occasions under Model M_t in closed population capture-recapture models, where $t = k$ (Otis *et al.* 1978). The general likelihood function for k independent list frames, $B_1, B_2, ..., B_k$, has the form

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N \mid N_{b_1}, ..., N_{b_1 ... b_k}) =$$

$$\binom{N}{N_{b_1}, ..., N_{b_1 ... b_k}} \prod_{l=1}^{k} p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}}, \qquad (5)$$

which has exactly the same structure as the likelihood introduced by Darroch (1958) and is discussed in great detail by Otis *et al.* (1978) and Seber (1982). The form of the estimated frame inclusion probabilities is

$$\hat{p}_{B_l} = \frac{N_{B_l}}{\hat{N}}, \quad l = 1, ..., k. \qquad (6)$$

Values of $\hat{N}$ are obtained by numerically solving the $(k - 1)$ degree polynomial in $\hat{N}$ resulting from the equality

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 ... b_k})} *$$

$$(1 - \hat{p}_{B_1}) \cdots (1 - \hat{p}_{B_k}) = 1. \qquad (7)$$

We then select as $\hat{N}$ as the root that maximizes the value of the likelihood function (5). Substituting this root into (6) yields MLEs of the k frame inclusion probabilities.

2.2 Population Total Estimation

Suppose the measured y_i values are available for all units on the k independent list frames. The estimated probability that the first element is included on at least one of the k list frames is

$$\hat{\pi}_1 = \hat{P}\left[\cup_{l=1}^{k} B_l\right] = 1 - (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}),$$

where $\hat{p}_{B_l} = N_{B_l} / \hat{N}$ and $\hat{N}$ is the MLE of N obtained from (7). From equation (7),

$$\frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 ... b_k})} (1 - \hat{\pi}_1) = 1$$

which simplifies to

$$\hat{\pi}_1 = \frac{N_{b_1} + \cdots + N_{b_1 ... b_k}}{\hat{N}}.$$

An estimated Horvitz and Thompson (1952) estimator of the population total is

$$\hat{\hat{Y}}_{\text{H-T}} = \frac{1}{\hat{\pi}_1} \sum_{i \in B_1 \cup ... \cup B_k} y_i$$

$$= \frac{\hat{N}}{N_{b_1} + \cdots + N_{b_1 ... b_k}} \sum_{i \in B_1 \cup ... \cup B_k} y_i = \hat{N} \bar{Y}_L,$$

where $\bar{Y}_L$ is the mean of distinct elements on the list frames. Thus, for k independent list frames, the estimated Horvitz-Thompson estimator coincides with the population total estimator proposed by Pollock, Turner and Brown (1994).

In some situations, values of the variable of interest, y_i, are not available for all units on the list frames. If the list frames are large in size, random samples are selected from each list frame and data are collected on those subsampled elements. If there are k list frames, it is possible to define 2^k domains. We consider an extension of Lund's (1968) estimator for the total of all units on the list frames,

$$\hat{Y}_{L,L} = \sum_{l=1}^{2^k - 1} N_l \bar{y}_l,$$

which is a weighted sum of $2^k - 1$ domain means, $\bar{y}_l$. The weights are given by the domain sizes. Further, the population total estimator is

$$\hat{Y} = \hat{N} \frac{\hat{Y}_{L,L}}{\sum_{l=1}^{2^k - 1} N_l}.$$

3. MULTIPLE LISTS PLUS AN AREA FRAME

3.1 Population Size Estimation

Joining multiple, individual list frames with an area frame sample is a solution to overcoming list frame deficiencies. Assume that the geographical area of interest is subdivided into U_A segments. Also, assume that a simple random sample of u_A segments is selected from U_A segments that cover the entire population. Therefore, the probability of a segment being selected is $p_A = u_A/U_A$. In some surveys, it is possible to subdivide the region into approximately equally-sized segments. In such cases the segment selection probability corresponds approximately to the proportion of area sampled. The inclusion of an area frame provides completeness of the target population (Hartley 1962). We assume that each reporting unit belongs to exactly one segment. Once a segment is selected, all reporting units within the segment are observed. For example, when estimating the number of bald eagle nests, each nest belongs to one and only one segment. However, this assumption is not always valid. Consider the case where a hog farm crosses segment boundaries. In this case, population elements may be associated with more than one segment. To address this problem, association rules linking population elements to segments are established at the estimation stage. See Faulkenberry and Garoui (1991) for more detail. The National Agricultural Statistics Service implements three correspondence rules that map elements in the population to sampled segments. The open, closed, and weighted segment estimators are described in Nealon (1984). Another related reference is Sirken (1970).

Consider the case of k independent list frames plus an area frame. The population size, N, and the list frame inclusion probabilities, p_{B_i}, $i = 1, ..., k$, are unknown parameters. The area frame inclusion probability $p_A = u_A/U_A$ is known. The likelihood function has the form

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N \,|\, p_A, n_a, n_{ab_1}, ..., n_{ab_1...bk}, N_{b_1}, ..., N_{b_1...b_k})$$

$$= \binom{N}{n_a, n_{ab_1}, ..., n_{ab_1...b_k}, N_{b_1}, ..., N_{b_1...b_k}} p_A^{n_A}(1 - p_A)^{N-n_A}$$

$$\prod_{l=1}^{k} p_{B_l}^{N_{B_l}}(1 - p_{B_l})^{N-N_{B_l}},$$

where n_A is the total number of elements in the u_A sampled area segments and n_a is the number of elements in the u_A sampled area segments which do not belong to any list frames. Similarly, $n_{ab_1}, ..., n_{ab_1...b_k}, N_{b_1}, ..., N_{b_1...b_k}$ are defined as the sizes of different domains. It is important to emphasize that the inclusion of an area frame may cause the value of N_{b_1} to change. N_{b_1} now corresponds to the number of elements on list frame B_1 which are not in the u_A selected area segments and not on any other list frame.

The MLEs of the parameters are given by $\hat{p}_{B_i} = N_{B_i}/\hat{N}$, where $\hat{N}$ is a solution to the k-th degree polynomial

$$\hat{N}(1 - p_A)(1 - \hat{p}_{B_1}) ... (1 - \hat{p}_{B_k}) =$$

$$(\hat{N} - n_a - n_{ab_1} - \cdots - n_{ab_1...b_k} - N_{b_1} - \cdots - N_{b_1...b_k}). \quad (8)$$

Numerical methods are essential for solving (8) for the MLE $\hat{N}$ of N. Among the k roots of (8), we select $\hat{N}$ that maximizes the likelihood.

Applying this methodology to one list frame and one area frame, we obtain

$$\hat{N} = N_{B_1} + \frac{n_a}{p_A}. \quad (9)$$

This estimator is also known as the screening estimator (Kott and Vogel 1995). The screening estimator categorizes elements into two distinct groups. The first group contains elements which belong to both the list and area frames and is called the overlap domain. Since it is assumed that all elements on a list frame belong to the area frame, the size of the overlap domain coincides with the number of elements on frame B_1 and has the value N_{B_1}. The second group contains elements in the area frame not included on the list frame(s) and is referred to as the nonoverlap domain. The size of the nonoverlap domain is an unobserved random quantity, N_a. The term n_a is the number of elements found in the u_A area segments which are not included on the list frame(s) following a specific association rule. An estimated value of N_a is n_a/p_A. Hence, an estimate of the population size is given by $\hat{N}$ in (9). The resulting MLE of p_{B_1} is

$$\hat{p}_{B_1} = \frac{N_{B_1}}{N_{B_1} + \frac{n_a}{p_A}}.$$

When multiple list frames are available, it is possible to combine them into a single list frame and use the above estimator to obtain an estimate of N. That is, consider the screening estimator

$$\hat{N}_2 = \hat{N} = N_{B_1 \cup ... \cup B_k} + \frac{n_a}{p_A} = N_{b_1} + \cdots + N_{b_k} +$$

$$N_{b_1 b_2} + \cdots + N_{b_1 ... b_k} + \frac{n_a}{p_A}. \quad (10)$$

Note that the screening estimator $\hat{N}_2$ is appropriate even when the list frames are *not* independent of each other. We discuss this further in Section 4.

Using this methodology for one area and two independent list frames yields the likelihood

$$\mathcal{L}(p_{B_1}, p_{B_2}, N \,|\, p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}) =$$

$$\binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}} p_A^{n_A} p_{B_1}^{N_{B_1}} P_{B_2}^{N_{B_2}}$$

$$(1 - p_A)^{N-n_A}(1 - p_{B_1})^{N-N_{B_1}}(1 - p_{B_2})^{N-N_{B_2}}.$$

The MLE of N is

$$\hat{N}_3 = \hat{N} = (2p_A)^{-1} * \left[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1b_2} - n_{ab_1b_2})\right] + (2p_A)^{-1} \sqrt{\left[(N_{B_1}+N_{B_2})p_A+(n_a-N_{b_1b_2}-n_{ab_1b_2})\right]^2 + 4p_A(1-p_A)N_{B_1}N_{B_2}}, \quad (11)$$

where $n_{ab_1b_2}$ denotes the number of elements included in the u_A sampled area segments that belong to both list frames. An estimate of the variance of $\hat{N}_3$ may be obtained using the Taylor series approximation of (11) and the asymptotic distribution of $(N_{B_1}, N_{B_2}, n_a, N_{b_1b_2}, n_{ab_1b_2})$.

3.2 Population Total Estimation

When y_i's are available for all elements on k independent list frames and for a sample of segments from an area frame, we consider an estimated Horvitz-Thompson estimator to estimate the population total. Recall that we assume the following:

1. The probability that a unit is included on the i-th list frame, p_{B_i}, is the same for all units.
2. The event that a unit is included on one frame is independent of its inclusion on another frame.
3. The probability that a unit is included in the area frame sample of u_A segments is $p_A = u_A/U_A$.

Since we consider the case where population units belong to exactly one area segment and all units within a sampled segment are observed, the third assumption is valid. Hence, the probability the i-th element is on at least one of the k list frames and/or the area frame sample is

$$\tilde{\pi}_1 = 1 - (1 - p_A)(1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}) = \frac{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}}{\hat{N}}.$$

The estimated Horvitz-Thompson population total estimator is

$$\hat{\hat{Y}}_{\text{H-T}} = \frac{\hat{N}}{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}} \sum_{i \in \text{sample}} y_i = \hat{N} \bar{y}_L,$$

where $\bar{y}_L$ is the mean of the distinct elements on list frames $B_1, \ldots, B_k$ and the elements in the area frame sample.

We can also use the screening estimator to estimate the population total. The known overlap domain total is combined with an estimator of the nonoverlap domain (NOL) total to yield $\hat{Y}_S = Y_L + \sum_{i \in \text{NOL}} y_i/p_A$. The NOL domain consists of elements on the area frame that are not on any of the list frames and $Y_L = Y_{B_1 \cup \ldots \cup B_K}$ is the total of the distinct units on the k list frames. In the subsampling case, we may replace Y_L in $\hat{Y}_S$ by Lund's estimator, given by

$$\hat{Y}_{L,L} = N_{b_1} \bar{y}_{b_1} + \cdots + N_{b_k} \bar{y}_{b_k} + N_{b_1b_2} \bar{y}_{b_1b_2} + \cdots + N_{b_1 \ldots b_k} \bar{y}_{b_1 \ldots b_k}.$$

4. DEPENDENT LIST FRAMES

We now consider the case where dependencies exist among list frames but where area and list frames remain independent. In capture-recapture experiments, for example, the probability an animal is captured on the second sampling occasion may depend on whether it was captured on the first sampling occasion. See Fienberg (1972), Cormack (1989), Wolter (1990), Pollock, Hines, and Nichols (1984), Huggins (1989), and Alho (1990) for specific examples.

We consider the case where we have two list frames, B_1 and B_2, that are dependent. Let p_{11} denote the probability of being included on both list frames. If B_1 and B_2 are independent, then $p_{11} = p_{B_1} p_{B_2}$ where p_{B_1} and p_{B_2} are inclusion probabilities for B_1 and B_2, respectively. Define $p_{10}(p_{01})$ as the probability of being included on frame $B_1(B_2)$ but not on frame $B_2(B_1)$. The probability of exclusion from both list frames is denoted by $p_{00} = 1 - p_{B_1} - p_{B_2} + p_{11}$.

The likelihood function is given by

$$\mathfrak{L}(p_{B_1}, p_{B_2}, p_{11}, N \mid p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}) = \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}} p_A^{n_A} (1 - p_A)^{N - n_A} (p_{B_1} - p_{11})^{N_{b_1} + n_{ab_1}} (p_{B_2} - p_{11})^{N_{b_2} + n_{ab_2}} p_{11}^{N_{b_1b_2} + n_{ab_1b_2}} (1 - p_{B_1} - p_{B_2} + p_{11})^{N - N_{b_1} - N_{b_2} - n_{ab_1} - n_{ab_2} - N_{b_1b_2} - n_{ab_1b_2}}. \quad (12)$$

Maximizing (12) with respect to p_{B_1}, p_{B_2}, p_{11} and N leads to the approximate solution

$$\hat{N} = N_{b_1} + N_{b_2} + n_{ab_1} + n_{ab_2} + N_{b_1b_2} + n_{ab_1b_2} + \frac{n_a}{p_A},$$

which coincides with the screening estimator $\hat{N}_2$. That is, $\hat{N}$ is also the estimator that is obtained by pooling the two list frames into a single list frame where the duplications are eliminated and the nonoverlap domain size is estimated using the area frame sample. Also, it can be shown that the two-stage maximum likelihood procedure of Sanathanan (1972) leads to:

$$\hat{N} = \frac{n_a + N_{B_1 \cup B_2}}{p_A + (1 - p_A)\dfrac{N_{B_1 \cup B_2}}{\hat{N}_2}}$$
$$= \hat{N}_2.$$

Thus, the maximum likelihood estimator and Sanathanan's estimator both coincide with the screening estimator. If information from two dependent list frames is available and the nature of the dependency is unknown, then we cannot estimate the individual parameters. When information from an independent area frame is available, all parameters are estimable. However, for estimating N, $N_{B_1 \cup B_2}$ is sufficient and no additional information is gained from N_{B_1}, N_{B_2}, and $N_{b_1 b_2}$.

Methods are available for modeling the dependence among k list frames when estimating population size and totals. Additional population information or information from an independent area frame is needed to accurately model the dependence. Fienberg (1972) and Cormack (1989) consider constrained log-linear models to model the dependence. On the other hand, Wolter (1990) uses external constraints such as a known sex ratio to estimate the population size in the dependence case. Another technique used is to model the inclusion probabilities as a function of the covariates. Alho, Mulry, Wurdeman and Kim (1993) use a conditional logistic regression model to estimate the probability of being enumerated in a census and apply the model to the 1990 Post-Enumeration Survey. The role of auxiliary variables in capture-recapture experiments with unequal capture probabilities is addressed in Pollock *et al.* (1984), Huggins (1989), and Alho (1990).

5. SIMULATION STUDY

We conduct a simulation study to assess the overall efficiency of different population size estimators for the special case of two list frames plus an area frame. This is the most feasible combination of sampling frames for real survey problems.

5.1 Design of the Study

In order to study both dependent and independent cases, we define the parameter θ that reflects the dependence structure between list frames B_1 and B_2. It has the same form as the odds ratio and is written formally as

$$\theta = \frac{p_{00} p_{11}}{p_{01} p_{10}}.$$

In the case of two list frames, the value of θ determines a unique solution for p_{11}. Our study varies the following factors:

Factor	Levels	Definition
N	500, 5000	Population size
p_A	0.05, 0.10, 0.20	Inclusion probability for area frame A
$p_{B_1}(=p_{B_2})$	0.7, 0.9	Inclusion probability for list frame $B_1 (B_2)$
θ	0.5, 1.0, 1.5, 2.0	Odds ratio

For each parametric combination, we generate data (n_a, $N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}$). One thousand Monte Carlo replications are generated for each parametric combination.

5.2 Estimators

We compare four population size estimators, $\hat{N}_1$, $\hat{N}_2$, $\hat{N}_3$, and $\hat{N}_4$. $\hat{N}_1$ is the Lincoln-Petersen estimator which does not incorporate area frame information. The estimator $\hat{N}_1$ is suitable when the list frames are independent. Since the estimator ignores information from the area frame sample, it is expected to be inefficient when information from an area frame is available. The screening estimator, $\hat{N}_2$, sums the overlap and nonoverlap domain estimates and is particularly suitable for the dependent list frame case. The third estimator, $\hat{N}_3$, is derived from the full, independent sampling frame likelihood function. This estimator exploits the information contained in the area and list frames and the fact that the list frames are independent ($\theta = 1$).

We expect $\hat{N}_3$ to be the best estimator when list frames B_1 and B_2 are independent whereas we expect $\hat{N}_2$ to be the best estimator in the dependent case. As a result, we also consider a pre-test estimator that tests for independence of the list frames. We define $\hat{N}_4$ to be $\hat{N}_2$ if there is strong evidence to believe that frames B_1 and B_2 are not independent. Otherwise, we take $\hat{N}_4 = \hat{N}_3$. Formally,

$$\hat{N}_4 = \begin{cases} \hat{N}_2 \text{ if GOF} > \chi^2_{1,0.05} = 3.84 \\ \hat{N}_3 \text{ otherwise,} \end{cases}$$

where GOF is the chi-square goodness-of-fit test statistic for testing H_0: $\theta = 1$ and is derived from the following two-way table.

	In B_1	Not In B_1	
In B_2	$n_{ab_1 b_2}$	n_{ab_2}	$n_{A \cap B_2}$
Not In B_2	n_{ab_1}	n_a	$n_{A \cap B_2'}$
	$n_{A \cap B_1}$	$n_{A \cap B_1'}$	n_A

Figure 1. Classification of Sampled Area Frame Elements

Figure 1 categorizes the n_A elements according to their presence on or absence from list frames B_1 and B_2.

5.3 Comparing the Estimators

Tables 2 and 3 display the percent relative bias and the percent relative root mean square error of the estimates $\hat{N}_1, \hat{N}_2, \hat{N}_3$, and $\hat{N}_4$ for population sizes of 500 and 5000, respectively. We scale the bias and the root mean square error by N in order to directly compare estimators based on different population sizes. A comparison of $\hat{N}_1$ with $\hat{N}_3$ shows the benefit of drawing an area frame sample. In practice, these benefits depend on the relative cost of the area frame sample. In this study, we do not take sampling costs into account. The probability of being included on both list frames, p_{11}, is given in parentheses in the θ column. When $p_B = p_C = .9, p_{11}$ must lie between .8 and .9. However, for θ ranging from .5 to 2, p_{11} varied only from .806 to .817.

The estimator $\hat{N}_2$ is unbiased for N and has the smallest percent relative bias. The estimators $\hat{N}_1$ and $\hat{N}_3$ are asymptotically consistent for N and yield biases close to 0 when $\theta = 1$. On the other hand, $\hat{N}_1$ and $\hat{N}_3$ have large biases when $\theta \neq 1$. The percent relative bias of $\hat{N}_4$ is smaller than that of $\hat{N}_3$ but it is not close to zero. The bias does not change significantly as p_A increases from .05 to .10 to .20.

When $N = 500$ and $p_B = p_C = .9$, $\hat{N}_3$ has the smallest percent relative root mean square error (% RRMSE). This is partly due to the fact that the limited range of p_{11} values is similar to the p_{11} value for the independence case (.810). The % RRMSE for $\hat{N}_3$ is 40 - 50 % smaller than that of $\hat{N}_2$. On the other hand, the % RRMSE of $\hat{N}_3$ is only 15 - 30 % smaller than that of $\hat{N}_1$. Therefore, when the list frames have very high inclusion probabilities, both $\hat{N}_1$ and $\hat{N}_3$ are much better than $\hat{N}_2$. Additionally, if area frame sampling costs are high, $\hat{N}_1$ may be a reasonable alternative estimator to $\hat{N}_3$. When $N = 500$ and$p_B = p_C = .7$, $\hat{N}_3$ has the smallest % RRMSE for the independence case. When $\theta = 2$, $\hat{N}_2$ has the smallest % RRMSE. If $N = 5000$ and $p_B = .7$, $\hat{N}_3$ has the smallest % RRMSE for only $\theta = 1$. For all other θ values, $\hat{N}_2$ yields the smallest % RRMSE. In all cases, $\hat{N}_3$ has very small variance and most of the % RRMSE is due to the bias in $\hat{N}_3$. For $\theta < 1$, $\hat{N}_3$ tends to have positive bias while for $\theta > 1$, $\hat{N}_3$ has negative bias. For the case of $N =$ 5000 and $p_B = .9$, $\hat{N}_3$ has the smallest % RRMSE for $\theta = 1$. $\hat{N}_2$ has the smallest % RRMSE for $\theta = .5$ and 2. For $\theta = 1.5$, there is no best estimator with respect to % RRMSE.

As expected, the percent relative root mean square errors of $\hat{N}_2$, $\hat{N}_3$, and $\hat{N}_4$ decrease as the value of p_A increases. Thus, as the area frame information increases, the % RRMSE decreases. Also, as the population size increases from 500 to 5000, the % RRMSE decreases. Since the values of p_A in our simulation are small, $\hat{N}_2$ has a large variance. On the other hand, even though $\hat{N}_3$ is biased, it has a very small standard error and results in a smaller % RRMSE. The estimator $\hat{N}_4$ reduces the bias of $\hat{N}_3$ but has a large standard error. Hence, $\hat{N}_4$ is not a particularly beneficial estimator. For larger values of θ and p_A, we expect $\hat{N}_2$ to perform better than $\hat{N}_3$. For the values of θ and p_A we considered, we recommend $\hat{N}_3$ over other estimators.

The value of % RRMSE for $\hat{N}_4$ is between that of $\hat{N}_2$ and $\hat{N}_3$ in most cases. We write the estimator $\hat{N}_4$ as $\hat{N}_4 = \delta\hat{N}_2 + (1 - \delta)\hat{N}_3$, where $\delta = 0$ or 1 based on the results of the goodness-of-fit test. The % RRMSE and % RBias of $\hat{N}_4$ need not be between those of $\hat{N}_2$ and $\hat{N}_3$ because δ is not independent of $\hat{N}_2$ and $\hat{N}_3$.

5.4 Limitations of the Study

The goal of our study is to compare the bias, standard error, and mean square error of four population size estimators. We assume that inclusion probabilities for both list frames are identical. Future studies may include unequal inclusion probabilities as well as larger values of θ. Clearly the benefit of $\hat{N}_3$ over $\hat{N}_1$ depends on the cost of sampling from an area frame. Our paper considers only small values of p_A. Small p_A values are associated with a high area frame sampling cost. Even in this case, we observe a significant reduction in % RRMSE and % RBias, thereby justifying the use of $\hat{N}_3$ over $\hat{N}_1$. We do not consider an objective function which incorporates sampling costs, % RRMSE, and % RBias.

Throughout this paper, we assume that all units have the same probability of being included on a given list frame. Haines (1997) considers the case where the inclusion probabilities are modeled as a function of a covariate. When inclusion probabilities are heterogeneous, larger units may have a higher list frame inclusion probability than smaller units. Heterogeneous inclusion probabilities play an important role in estimating population totals when the response variable has a highly skewed distribution or has rare values. Haines (1997) also presents two stratification procedures that are useful when area and list frames are stratified on the same variable. These results will be presented in future publications.

6. DISCUSSION

The primary focus of this paper is population size estimation based on several sampling frames. Information from area and/or list frame(s) is collected and combined to obtain various estimators. We derive population size estimators when information is available only on k independent list frames and also when information is available on an area frame sample in addition to the list frames. We conduct a simulation study to compare the performance of the estimators in the special case of two list frames plus an area frame. Based on our simulation study, we recommend the estimator derived from the full, independent likelihood, $\hat{N}_3$, for the case where the list

Table 2
Simulation Results for N = 500

			p_A					
			.05		.10		.20	
p_B	θ		% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	$\hat{N}_1$	62.30	66.01	60.64	64.04	63.26	66.81
		$\hat{N}_2$	0.30	49.07	-0.75	32.37	0.85	22.58
		$\hat{N}_3$	55.52	58.95	48.15	51.15	40.53	43.32
		$\hat{N}_4$	48.15	58.88	37.88	49.25	24.95	38.80
	1 (.490)	$\hat{N}_1$	0.47	19.26	1.01	19.08	-0.11	19.45
		$\hat{N}_2$	0.45	57.34	0.34	39.61	0.88	27.25
		$\hat{N}_3$	0.43	18.21	0.83	16.93	0.14	15.75
		$\hat{N}_4$	2.40	27.57	1.39	22.94	0.29	17.96
	1.5 (.508)	$\hat{N}_1$	-35.60	40.06	-36.48	40.58	-35.69	40.26
		$\hat{N}_2$	3.11	66.43	-5.08	41.96	0.30	28.79
		$\hat{N}_3$	-32.07	36.79	-31.01	35.28	-24.04	28.88
		$\hat{N}_4$	-22.74	47.62	-26.21	37.57	-17.06	30.38
	2 (.522)	$\hat{N}_1$	-60.07	62.91	-61.31	64.06	-60.41	63.28
		$\hat{N}_2$	-6.12	66.59	-1.15	46.68	1.67	30.99
		$\hat{N}_3$	-55.36	58.35	-51.21	54.19	-40.89	43.99
		$\hat{N}_4$	-41.39	63.79	-34.79	55.45	-18.60	41.35
.9	.5 (.806)	$\hat{N}_1$	5.37	6.79	5.27	6.63	5.59	6.97
		$\hat{N}_2$	0.08	14.78	-0.06	10.17	-0.06	6.55
		$\hat{N}_3$	5.04	6.44	4.62	5.93	4.24	5.53
		$\hat{N}_4$	5.94	9.48	5.03	7.05	4.34	5.72
	1 (.810)	$\hat{N}_1$	0.30	5.01	0.17	5.01	0.25	4.94
		$\hat{N}_2$	0.78	20.72	0.41	14.06	-0.06	9.03
		$\hat{N}_3$	0.33	4.83	0.20	4.68	0.17	4.24
		$\hat{N}_4$	3.23	13.79	1.88	9.35	1.00	5.98
	1.5 (.814)	$\hat{N}_1$	-4.29	7.07	-4.39	7.32	-4.55	7.37
		$\hat{N}_2$	-0.65	21.52	0.35	15.88	0.002	10.27
		$\hat{N}_3$	-4.07	6.78	-3.83	6.73	-3.49	6.15
		$\hat{N}_4$	-0.43	13.77	-1.18	10.92	-1.43	8.20
	2 (.817)	$\hat{N}_1$	-8.28	10.27	-8.40	10.36	-8.33	10.32
		$\hat{N}_2$	-0.29	25.59	0.39	17.66	0.35	11.41
		$\hat{N}_3$	-7.80	9.82	-7.35	9.38	-6.30	8.20
		$\hat{N}_4$	-2.52	17.96	-3.10	14.02	-2.73	10.33

Table 3
Simulation Results for N = 5000

			p_A					
			:05		.10		.20	
p_B	θ		% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	$\hat{N}_1$	61.47	61.82	61.39	61.76	61.69	62.04
		$\hat{N}_2$	-0.18	15.78	0.26	10.65	-0.15	6.72
		$\hat{N}_3$	54.84	55.17	49.06	49.38	39.38	39.65
		$\hat{N}_4$	19.73	38.12	4.77	19.52	-0.01	7.21
	1 (.490)	$\hat{N}_1$	-0.28	6.14	-0.13	5.99	0.35	6.15
		$\hat{N}_2$	0.43	18.14	0.47	12.85	-0.20	8.34
		$\hat{N}_3$	-0.22	5.82	-0.03	5.35	0.16	4.88
		$\hat{N}_4$	0.26	9.82	-0.04	7.44	0.11	5.95
	1.5 (.508)	$\hat{N}_1$	-36.21	36.68	-36.29	36.78	-35.90	36.38
		$\hat{N}_2$	0.41	20.39	-0.16	14.21	0.39	9.55
		$\hat{N}_3$	-32.87	33.37	-29.97	30.49	-24.13	24.66
		$\hat{N}_4$	-19.11	31.15	-11.51	23.92	-3.12	14.03
	2 (.522)	$\hat{N}_1$	-61.04	61.3	-60.53	60.81	-60.64	60.92
		$\hat{N}_2$	0.40	20.09	0.60	15.43	0.31	9.67
		$\hat{N}_3$	-55.69	55.96	-50.24	50.55	-41.46	41.76
		$\hat{N}_4$	-14.10	36.31	-2.34	20.96	0.26	9.84
.9	0.5 (.806)	$\hat{N}_1$	5.56	5.70	5.52	5.67	5.54	5.68
		$\hat{N}_2$	-0.12	4.55	0.11	3.19	-0.03	2.08
		$\hat{N}_3$	5.21	5.35	4.86	5.01	4.22	4.35
		$\hat{N}_4$	4.97	5.41	3.64	4.88	2.26	3.79
	1 (.810)	$\hat{N}_1$	-0.02	1.58	0.08	1.55	0.01	1.57
		$\hat{N}_2$	-0.09	6.16	-0.17	4.08	-0.14	2.79
		$\hat{N}_3$	-0.03	1.53	0.05	1.48	-0.02	1.35
		$\hat{N}_4$	0.37	3.19	0.11	2.18	0.09	1.89
	1.5 (.814)	$\hat{N}_1$	-4.66	5.00	-4.52	4.85	-4.61	4.90
		$\hat{N}_2$	-0.25	7.54	0.11	4.95	-0.09	3.14
		$\hat{N}_3$	-4.39	4.73	-3.96	4.32	-3.55	3.85
		$\hat{N}_4$	-2.50	6.31	-2.26	5.02	-1.84	3.82
	2 (.817)	$\hat{N}_1$	-8.45	8.68	-8.38	8.60	-8.46	8.69
		$\hat{N}_2$	-0.21	7.86	-0.06	5.29	0.01	3.73
		$\hat{N}_3$	-7.95	8.18	-7.39	7.61	-6.49	6.73
		$\hat{N}_4$	-3.76	8.80	-2.77	6.99	-1.25	4.97

frames are independent or nearly independent. For the moderate to strong dependence cases, we recommend the screening estimator, $\hat{N}_2$.

We also study population total estimation. We consider two scenarios for estimating population totals. In the first case, we assume that observations are available on all units that comprise the list frames. In contrast, the second case assumes that information is available only on subsamples from each of the list frames. We consider an estimated Horvitz-Thompson estimator if list frames are independent and a screening estimator to estimate the population total if the list frames are dependent.

In this paper, our focus is on population size estimation. In practice, one may be interested in estimating population totals for several characteristics based on multi-stage samples involving unequal inclusion probabilities. Relevant papers on this topic include Bankier (1986), Skinner (1991), and Skinner, Holmes, and Holt (1994).

7. ACKNOWLEDGEMENTS

The authors thank the editor and two referees for useful comments on an earlier version of the article. This research was partially funded by the U.S. Geological Survey, Biological Resources Division. Christine Bunck is the BEST Program Manager. 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.

ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.

DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.

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. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

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

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

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

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.

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, DC.

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.

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.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43, 1, 142-152.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (2nd Edition). New York: Macmillan.

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.

SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

SKINNER, C.J., HOLMES, D.J., and HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 333-347.

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

Survey Methodology, June 1998
Vol. 24, No. 1, pp. 89-98
Statistics Canada

Temporary Mobility and Reporting of Usual Residence

NANCY BATES and ELEANOR R. GERBER[1]

ABSTRACT

Temporary mobility is hypothesized to contribute toward within-household coverage error since it may affect an individual's determination of "usual residence" – a concept commonly applied when listing persons as part of a household-based survey or census. This paper explores a typology of temporary mobility patterns and how they relate to the identification of usual residence. Temporary mobility is defined by the pattern of movement away from, but usually back to a single residence over a two-three month reference period. The typology is constructed using two dimensions: the variety of places visited and the frequency of visits made. Using data from the U.S. Living Situation Survey (LSS) conducted in 1993, four types of temporary mobility patterns are identified. In particular, two groups exhibiting patterns of repeat visit behavior were found to contain more of the types of people who tend to be missed during censuses and surveys. Log-linear modeling indicates that temporary mobility patterns are a significant predictor of usual residence, even when controlling for the amount of time spent away and demographic characteristics.

KEY WORDS: Temporary mobility; Usual residence; Household rosters; Coverage.

1. INTRODUCTION

The fundamental challenge in any census of population is the accurate and complete count of every person within that population. Consequently, the extent to which people are missed or undercounted during a census is arguably the most important measure by which it is evaluated. Most censuses and household-based surveys begin with a roster question designed to list all "usual residents" of a household.

Research evaluating the quality of census data suggests that coverage error is a problem. In 1990, the U.S. Post Enumeration Survey (PES) and demographic analyses estimated that the net national undercount was approximately 2% (Hogan 1993; Robinson, Ahmed, Das Gupta and Woodrow 1993). Other research suggests that coverage error in current surveys (such as the U.S. Current Population Survey) is even larger than undercoverage occurring during decennial censuses (Shapiro, Diffendal, Cantor 1993; Chakrabarty 1992; Pennie 1990; Hainer, Hines, Martin and Shapiro 1988). Research by Fein and West (1988) and Shapiro *et al.* (1993) suggest that failure to count all persons within a housing unit is a larger component of total coverage error than failure to count persons as a result of missing a housing unit. Others report that within-household omissions account for about one-third of all census omissions (Ellis 1994; Fay 1989a).

Coverage research also indicates that persons who are undercounted are not randomly distributed among the population. For example, blacks and Hispanics are undercounted at a higher rate than non-Hispanic whites (4.6% and 4.0%, respectively, compared to 0.7%; Hogan 1993). Persons who reside in multi-unit structures (such as apartments) and those who rent are also more likely to be missed (Griffin and Moriarity 1992; Moriarity and Childers 1993; Ellis 1993).

This paper concentrates on a dimension long hypothesized to contribute to within-household coverage error. This dimension focuses on temporary mobility into and out of a residence over a period of time. Specifically, we examine movement in terms of the number of places a person may visit, the number of visits he/she makes and the amount of time he/she spends there. This analysis examines whether or not mobility may be a factor influencing coverage and indeed be a good indicator of household attachment. We hypothesize that a person's level of mobility tends to influence a household respondent's decision when defining that person as a usual resident and, consequently, someone he/she would or would not include on a census report.

2. BACKGROUND

The movement from one geographical location to another is usually signified by a change of address, movement of possessions and so on. This type of mobility is commonly referred to as geographic mobility. In addition to geographic mobility, there exists a more subtle form of mobility that is not so clearly defined – temporary mobility. Defined here, temporary mobility refers to the temporary and sometimes patterned movement away from a residence and encompasses both long and short, frequent and infrequent overnight stays. This type of mobility has been described as "one of the key features of irregular and complex households" (de la Puente 1993). One example of this is found in Haitian immigrant communities where typical household structure consists of a relatively

[1] Nancy Bates, Office of the Director, U.S. Bureau of the Census, Room 2031, Federal Building 3, Washington, DC 20233, and Eleanor R. Gerber, Center for Survey Methods Research, U.S. Bureau of the Census, Room 3133, Federal Building 4, Washington, DC 20233 U.S.A.

permanent "nuclear core" and a more mobile "fluid periphery." The fluid periphery consists of related and nonrelated newcomers, staying for short periods of time, and members of the household who visit Haiti on a regular basis and can be away weeks or months at a time (Wingerd 1992).

Temporary mobility is not limited to special communities. Many examples can be found in the wider community, including mobility associated with long term business or vacation travel, attendance at college, custody situations, and persons who maintain a presence in one or more households over a given period of time. This mobility in the fluid periphery, or temporary mobility, differs from geographic mobility because it consists of movements away from, but usually back to, a single residence over time. Members of this fluid periphery present conceptual difficulties for respondents in identifying which members to include in a census or survey. Movement of these persons may not involve a permanent change in address, and thus can blur the concept of who is defined as living or staying at a given address.

Given that there is little literature on temporary mobility, studies on geographic mobility and household structure provide a good starting point for forming our hypotheses about temporary mobility. According to the March 1994 Current Population Survey, young adults between 20-24 are reported to have the highest rates of geographic mobility, with one-third having moved between March 1993 and March 1994. Differences by race are also evident with a higher rate of mobility among blacks and Hispanics (19.6% and 22.4%, respectively) compared to whites (16.0%, see Hansen, 1994). Finally, tenure is also closely correlated with geographic mobility – renters were four times more likely than homeowners to have moved between 1993 and 1994. Obviously, these geographic movers share many of the same characteristics as some undercounted populations.

The kind of mobility with which we are concerned may also be a reflection of socioeconomic status. Temporary mobility, transitory situations, and peripheral connection to households can represent a means of adjusting for a lack of resources (Lipton and Estrada 1993). Hudgins and Holmes (1993) suggests that the undercounting of young black males is a result of their social and economic marginality evidenced in part by a lack of stable residences and relatively permanent mailing addresses. One facet of this may involve temporary movement to extended families or "kin" networks in order to receive family or financial assistance. This phenomenon of extended or kin networking among blacks has also been documented extensively by ethnographic studies (Martin and Martin 1985; Stack 1974; Hainer *et al.* 1988). These living arrangements suggest nontraditional (or at least non-nuclear) household formations which could contribute to coverage error, especially if a person participates in kin networks by moving back and forth among them.

Finally, Montoya (1992) describes a very different household composition that is characteristic of some recent Hispanic immigrant communities. Like kin-network households, they contain people who come and go, however, the members are "loosely tied, ephemeral, and alienated" and often composed of young migrant men who work and sleep in different shifts and have virtually no social ties with one another. Several other ethnographers have identified similar households in other Hispanic communities across the United States (Velasco 1992; Mahler 1993; Romero 1992.) They found that census coverage in such households was often restricted to those individuals who were actually present when the enumerator arrived.

3. METHODOLOGY

Data for this analysis come from the Living Situation Survey (LSS), a survey specifically designed to gather information about household membership, social attachments, mobility and the assignment of usual residence. The LSS was a voluntary survey conducted by the Research Triangle Institute (RTI) and sponsored by the U.S. Census Bureau between May and September of 1993. The sample was stratified to oversample for high and medium minority areas (*i.e.*, greater than 80% black or Hispanic, between 40% and 80% black or Hispanic) and areas containing renters (*i.e.*, greater than 40% renters). To increase the efficiency of the sample design, RTI used housing unit data previously collected from a multistage probability sample used in the 1992 National Household Survey on Drug Abuse (NHSDA).

The first portion of the LSS interview was conducted in-person with the most knowledgeable household respondent, in most cases, the householder (by U.S. Census Bureau definition, this refers to the person in whose name the house is owned or rented). These householders provided a roster and then answered demographic questions for themselves as well as all other listed persons. Through a series of 13 extensive roster probes, the questionnaire rostered "core" household residents but also included many persons having a less permanent presence. Persons with a more tenuous attachment were brought in by asking probes about who had spent the night there during the reference period, who was considered a household member even if they were staying elsewhere, and who considered the residence their permanent address or a place they received mail or phone messages (see Sweet 1994). (The length of the reference period varied depending upon the date of the interview. References periods began on the first day of the month two months prior to the interview month and ended on the day of the interview. Accordingly, interviews conducted toward the end of the month had a longer reference period than interviews conducted near the begining). In total, 999 households were interviewed nationwide. Using the broad rostering technique, a total of 3,549 people were listed.

The next step in the survey was to weed out rostered individuals determined to be only "casual visitors" to the household. Individuals were defined as casual visitors if: 1) their usual residence was considered by the householder to be someplace other than the sample housing unit *and* 2) they had stayed at the household for one week or less during the reference period. This screening process identified persons from the broad rostering technique who had only a casual attachment to the household. Of the 3,549 persons rostered, 712 were considered to be casual visitors. (Of the 712 casual visitors, 77% were related to the household respondent, 93% were non-Hispanic, 84% were white and 58% were female). For several reasons, casual visitors were ineligible for the remainder of the questionnaire. First, we assumed that casual visitors do not meet the Census Bureau definition of a usual resident at the interview household and second, excluding this group from the bulk of the questionnaire greatly reduced the time and resources required to carry out the survey.

After follow-up for converting refusals and other noninterviews, the final response rate for the household-level portion of the interview was 79.5%. (Follow-up actions included sending refusal conversion letters, having field supervisors call directly, make repeat visits, and re-assign interviewers. Respondents were contacted an average of 1.9 times; nonrespondents an average of 5.9 times). Considering the population, this was considered to be an acceptable rate of response. Nonetheless, since we suspect that nonresponse is highly related to coverage issues such as mobility, it is likely that this level of nonresponse has some effect upon our estimates. More discussion on this is included in the description of the individual questionnaire below.

The next part of the survey was a self-reported individual-level questionnaire. This part of the survey contained questions about temporary mobility as well as self-reported demographics. Respondents were asked if they had stayed overnight at any other place beside the interview household during the reference period. If so, interviewers used a calendar to record each place and the dates stayed. Interviewers also gathered information about the type of each place stayed, the individual's attachment to each place, and the reason(s) for going there.

Each of the householders answered the individual-level questionnaire for himself/herself. Additionally, all rostered persons who had stayed away for eight or more nights during the reference period answered the individual-level questionnaire. All persons identified as college students and persons with no usual residence were also eligible for an individual interview. Finally, the individual questionnaire was also given to a simple random 10% sample of LSS households. Within these households, individual interviews were attempted with each person on the roster, *with the exception of casual visitors.* This somewhat complex selection criterion resulted in a base of persons representing people with a greater-than-casual association to the interview households, all of whom are included in the analyses reported below (N = 1,451).

The individual-level portion of the questionnaire had a response rate of 85.3%. The majority of individual interviews were conducted in-person (96%) and most of the adult interviews (89%) were self-reported while all interviews with children were conducted by a knowledgable proxy. Because the householders answered basic living situation questions and demographic questions for *all* rostered individuals, we had some means for examining the characteristics of the approximately 15% who were selected for the individual questionnaire but did not respond. We found no significant sex or age differences between nonrespondents and respondents but we found that a disproportionate percentage of nonrespondents were black. We also found that nonrespondents were more likely to have spent more than one week away from the interview household than respondents. These findings shed some light on how representative our individual sample is both demographically and with respect to temporary mobility. Because nonrespondents were reported to be away more than respondents, we suspect the potential 'selectivity' bias may have underestimated our mobility measures.

Household and individual-level weights were applied to adjust for the oversampling, the selection criteria for the individual-level survey and for nonresponse (see Lynch, Witt, Branson and Ardini 1993). All analyses were conducted using Contingency Table Analysis for Complex Sample Designs (CPLX), a computer variance estimation program designed to adjust for the LSS's complex sample design effects (see Fay 1989b; 1985).

3.1 Typology of Temporary Mobility

The typology which we present is empirically based. That is, the particular groupings of visits and destinations was derived analytically and not theoretically. Therefore, the categories we identify do not represent groups of persons with identical characteristics or in identical circumstances. Rather the typology should be regarded as an attempt to represent the complex underlying reality involved in mobile living situations. It is our hypothesis that such mobility has an affect on the strength of the social tie between an individual and a particular household, and that these ties influence the judgment of the household respondent in deciding who is a usual resident of the household. Time away, number of visits and number of destinations are an indirect measure of the strength of such ties.

Our typology of temporary mobility was created using two dimensions of overnight movement outside the interview household. The first dimension taps into the variety of places a person visited over the reference period. This provides some idea of how many places other than the interview household that a person might have attachments to. The second dimension taps the frequency of movements outside the interview household by counting the number of times a person left for a period of one or more nights.

The use of these factors as a measure of the strength of attachment to a household is confirmed by ethnographic descriptions of highly mobile living situations. The pattern of movement represented in our typology reflects many different social processes, such as dispersed attachment to extended kin households (Stack 1974; Dressler, Hoeppner and Pitts 1985), immigration patterns (Wingerd 1992), and adaptation to poverty (Hainer 1987; Valentine and Valentine 1971).

The LSS included several exploratory open-ended questions designed to examine respondents perception of the reasons for their mobility. The questions asked the reasons for going and reasons for return for particular trips. We had hoped that these questions would provide us with a more direct assessment of the underlying social patterns that cause temporary mobility. Unfortunately the answers to these open ended questions were difficult to code without making unwarranted assumptions, largely as a result of the way in which they were expressed. As a result, we did not incorporate these reasons when formulating the typology.

Each "move" was defined as a stay made outside the interview household for at least one night. For example, if a person left to spend three days at a girlfriend's, then moved from there to a relative's for one night before returning to the interview household that person would be assigned as having two total places with two total visits (one visit apiece). Conversely, if a person left to stay overnight at a friend's then returned to the household and then two weeks later returned to the same friend's home for a second visit, that person would be assigned one place with two total visits (two repeat visits). The first example exemplifies a potential bias in this method, that of counting each unique place visited during one extended trip outside the interview household as an independent move (such as a vacation with multiple destinations). On the other hand, this method also captures the movement of "floaters" by counting each separate place visited during one move away from the household as a separate move.

A single mobility measure using various combinations of the number of places and number of moves was constructed. In all, five categories were created with efforts made to identify different patterns of movement by separating out those making repeat visits to the same places. Our first category depicts persons who stayed all nights of the reference period at the interview household and represents persons with no temporary mobility (the "Non-mobile"). The second category consists of persons who, according to the calendar, reported only one visit to one place (the "1-shots"). The "Boomerangs" reflect persons making repeat visits to one place only. The "No-repeats" are characterized as persons who traveled to more than one place, but never the same place twice. And finally, the "Floaters" stayed overnight at several different places, making repeat visits back to at least one of these places (see table 1).

Table 1
Temporary Mobility Typology

Number of Places Visited	Number of Visits				
	0	1	2	3	4
0	Non-mobile				
1		1-Shots	Boomerangs	Boomerangs	Boomerangs
2			No Repeats	Floaters	Floaters
3				No Repeats	Floaters
4					No Repeats

4. CHARACTERISTICS OF MOBILITY TYPES

Table 2 presents the weighted frequencies for the mobility typology. Slightly more than half of the persons administered the individual questionnaire reported no mobility outside the interview household during the reference period. The largest concentration of persons who were mobile fell into the 1-shot category, that is, they reported making only one move outside the interview household to one place (26%, overall). Eleven percent comprised the Boomerang category reporting a more repetitive pattern of two or more visits to a single place while 7% reported the less patterned, yet highly mobile "No repeat" category. The Floaters comprised the smallest group with 4%.

Table 2
Typology of temporary Mobility by Sex and Hard-To-Enumerate (HTE)* Status (Weighted % and standard errors)

MOBILITY TYPE	Total Weighted Percent (s.e. in paren.)	SEX		HTE STATUS	
		MALE	FEMALE	NON-HTE	HTE
Non-mobile	52% (14.0)	40% (13.7)	67% (13.6)	53% (14.3)	38% (7.8)
1-Shots	26% (10.4)	35% (13.9)	16% (7.0)	27% (10.6)	6% (2.9)
Boomerangs	11% (4.0)	15% (5.7)	6% (2.9)	10% (4.1)	21% (9.1)
No Repeats	7% (2.9)	6% (2.4)	8% (4.3)	7% (3.0)	6% (5.4)
Floaters	4% (1.0)	4% (1.3)	3% (1.3)	3% (0.9)	29% (9.9)
Unweighted N	1,451	653	798	1,375	76
Jackknife chi-square**		X^2 =2.03, p<.05, $d.f.$ =4		X^2 for distribution excluding non-mobile category =2.14, p < .05, $d.f.$ = 4	

* The hard-to-enumerate group includes black and Hispanic males aged 18-29.

**See Fay 1985 for documentation of Jackknife chi-square test for complex samples.

Tables 2 also illustrates selected demographics for the five mobility categories including gender breakouts which illustrate a higher mobility propensity for males than females. Approximately 60% of the males reported at least one visit outside the interview household, which was significantly higher than females at approximately 33%. This gender difference in temporary mobility is much more pronounced than in geographic mobility where the difference between the male and female move rate is only around 1% (17% of the male population moved between 1993 and 1994 compared to 16% for females, see Hansen 1994). This suggests that temporary mobility is more common than geographic mobility and that the demographic characteristics associated with it are different as well. Military travel could explain the gender differences in temporary mobility, as could travel for business with males having a higher active-duty/population ratio and employment/population ratio compared to females (U.S. Department of Labor 1994).

The right side of Table 2 integrates several demographic characteristics to create a subgroup known to have high rates of undercount in previous censuses. This group is comprised of males between 18 and 29 who are black or Hispanic. This subgroup is sometimes referred to as the "hard-to-enumerate" or HTE population. Only a small percentage of the LSS sample met the HTE criteria, but an examination of this group's mobility reveals very different patterns compared to the non-HTE group.

First, the HTE group appears more mobile to begin with – over 60% indicated spending at least one night someplace other than the interview household compared to less than 50% for non-HTEs. Second, the distribution of mobile categories differs significantly by HTE status. The majority of non-HTEs who are mobile are concentrated in the 1-shot category whereas the HTEs who are mobile are more concentrated in the repeat movement categories (Boomerangs and Floaters with 21% and 29%, respectively).

We also examined the distributions for temporary mobility by race (white, black, Hispanic, and other) and age (0-17, 18-29, 30-49, 50+). Overall, temporary mobility did not vary significantly by either, yet some interesting trends were noticeable. A relatively large concentration of Hispanics were found in the No-Repeat category (19%) and blacks in the Floater group (9%). A higher percentage of blacks were Non-mobile (66%) compared to whites (52%), in spite of the fact that blacks have higher rates of geographic mobility than whites. Finally, young adults between 18 and 29 appeared more mobile than other age groups (close to 70% of this age group spent at least one night away from the interview household) and a disproportionate percentage of this group were Floaters (14%). The lack of statistical significance among some of these trends may be an artifact of sample size. Alternatively, temporary mobility may be sufficiently different from geographic mobility such that it does not share the same characteristics of traditional 'movers'.

Another important variable hypothesized to correlate with the pattern of temporary mobility is the amount of time spent away on visits. The U.S. Census Bureau residence rules vary in the use of time as a criterion for usual residence. For example, persons who work in another city during the week but return home on weekends are to be counted at the place where they "live and sleep" the majority of the time – in this case, at the place they live during the week. However, a child living away at boarding school is to be counted at the parent's residence even though he/she probably spends the majority of time at the school. Likewise, a person staying at a group quarters on Census Day (*e.g.*, a college dorm or a jail) is counted at that place, regardless of their living situation the rest of the year. Gerber (1994) found that respondents also use time to varying degrees when defining household rosters – in certain situations, she found no clear relationship between being rostered and the amount of time spent at a place. Instead, things like household membership and relationship seemed to factor more heavily in the decision-making process.

Nonetheless, it makes intuitive sense that the amount of time spent away plays some part in the householder's determination of where to count someone. In order to see how our mobility categories varied in term of length of time spent away, the sum of the total number of nights spent away during all visits in the reference period was divided by the total number of nights in the reference period and then expressed as a percentage. Table 3 presents this time measure expressed in terms of being away more or less than half of the reference period.

Table 3
Time Spent Away from the Interview Household during the Reference Period (Weighted % and standard errors)

Away 50% of time or more?	1-Shots	Boomerangs	No Repeats	Floaters	Total
No	94% (4.4)	73% (11.5)	98% (1.4)	63% (10.3)	88% (3.6)
Yes	6% (4.4)	27% (11.5)	2% (1.4)	37% (10.3)	12% (3.6)
Unweighted N	314	186	101	134	735

Jackknife chi-square = 1.71, $p < .05$, $d.f.$ = 3

Both the Boomerangs and Floaters were more likely than other groups to spend half or more of the reference period someplace other than the interview household. This supports the notion that the repeat visit patterns underlying these two groups are associated with an increase in total time spent away. It also suggests a higher degree of residential ambiguity especially for the Floaters. Since members of this group report visits to at least two places in addition to the interview household, it is unclear whether those away more than half the time are spending a majority of time at any one place. If time spent at each place is roughly equal, it is easy to imagine Floaters not being rostered at any of them or at more than one of them. Conversely, by definition we can assume the Boomerangs who were away more than half the reference period spent the majority of their time at the only other place they reported visiting. Assuming time plays a role in defining a sense of household membership, then presumably, the Boomerangs have a better chance of being counted because the majority of their time is being spent at the other place.

5. USUAL RESIDENCE AND MOBILITY

We next explored whether temporary mobility has an impact on the household respondent's determination of a person as a "usual resident". On the 1990 U.S. census form, respondents were instructed to list persons at the place where the person lives or sleeps most of the time. The LSS asked household respondents whether they considered the interview household to be the "usual residence, that is the place where [you/NAME] live(s) and sleep(s) most of the time". They were also asked to report whether "[you/NAME] have a usual residence somewhere else?" While this method is not a perfect replication of a census roster it provides an approximation of who, out of all those rostered during the LSS, the householder might naturally have included or excluded on a census form or current survey.

Table 4 presents a cross-classification of usual residence assignment by mobility status. A combination of the usual residence questions resulted in four classification possibilities: usual residence at the interview household only, usual residence at someplace other than the interview household only, usual residence at both the interview household and another place, and usual residence at no place. (The category of "no place" was extremely small (less than 1%) and was combined into the category of "other place"). Assuming that answers of "other place" equate to being left off the census form, we see that overall, only around 4% of persons with a greater-than-casual association to the interview households might have been left off. Overall, the distribution of usual resident classifications significantly differed according to mobility type.

As might be expected, nearly all of the persons who spent every night at the interview household during the reference period were considered usual residents there (rounded to 100%). The most obvious deviation among categories is noticeable for the Boomerangs and Floaters. Between 20-25% of the people in these two groups were characterized by household respondents as usual residents someplace other than the interview household. This looks very different from both the 1-shots and No-repeat groups, where only 2% and 5%, respectively, were considered usual residents someplace else. These results suggest that the latter two groups typify mobility associated with pleasure or business but for persons with a firm tie to the household while the Boomerangs and the Floaters are more likely to include persons with a less-established association to the household. For this reason, and the fact that a sizable percentage of the HTE population were found in these two categories, the Boomerangs and Floaters arguably have the more interesting coverage implications and raise several questions. For example, do these persons get counted at one place, all places or no place? Additionally, where should they be counted?

Table 4
Where Does Household Respondent Consider Person to be a "Usual Resident" ? (Weighted % and standard errors)

Where Usual Resident ?	Non Mobile	1-Shots	Boomerangs	No Repeat	Floaters	Total
Interview HH Only	100% (0.2)	97% (2.0)	71% (12.1)	95% (4.2)	70% (10.0)	95% (1.7)
Some Other Place	0% (–)	2% (1.8)	25% (11.0)	5% (4.2)	20% (9.4)	4% (1.5)
Both Places	0% (–)	1% (0.4)	4% (2.1)	0% (–)	10% (7.3)	1% (0.5)
Unweighted N	716	314	186	101	134	1,451

Jackknife chi-square = 2.79, $p < .05$, $d.f.$ = 8

That a relatively large percentage of the Boomerangs and Floaters are considered residents some place other than the interview household suggests the potential for undercounting. On the other hand, 10% of the Floaters are defined as usual residents at both the interview household and another place suggests potential for overcoverage. The weighted number of Boomerangs and Floaters in these uncertain residency situations (usual residents elsewhere or at both places) represent approximately 4% of the total population. From this more global perspective, it seems that a non-trivial segment of the population is at risk of some type of coverage error.

6. MODELING OF USUAL RESIDENCE AND MOBILITY

Our final section statistically models the household respondent's determination of usual residence. This analysis goes beyond the descriptive findings of the typology to explore whether mobility impacts the householder's conceptualization of residence. The assignment of usual residence by the householder served as the dependent variable in a series of models. The dependent variable consisted of two categories: 1) usual resident at the interview household and 2) not a usual resident at the interview household. Persons considered to have a usual residence at both the interview household and another place were put into the first category. Predictor variables included age, sex, race, time away, and the mobility typology. The final models reported in Table 5, all of which include terms for the interaction of the independent variables, are equivilent to logit models for usual residence.

The first model tested mobility as a dichotomous measure: those with no mobility (the Non-mobile) and those having spent at least one night away from the interview household (the 1-shot, No-Repeat, Boomerang and Floater categories combined). This model established first whether temporary mobility was a significant predictor of residency status regardless of the mobility pattern exhibited. This "first-cut" was necessary because approximately 50% of the sample fell into the Non-mobile category and second, because the Non-mobile group was extremely skewed toward the usual resident category of the dependent variable. Consequently, models that attempted to include all five categories of the mobility typology were misspecified due to a large number of zero fitted cells.

Results from the model with the dichotomous mobility measure and sex yielded a relatively good "fit" of the data (Jackknife X^2 for overall goodness of fit = .28, *d.f.* = 2, $p = .27$. Neither race nor age improved the fit. Parameter estimates indicated that persons in the Non-mobile category were more likely to be classified as usual residents than those having some mobility (not shown).

Having established that mobility was significantly related to residency status, we next explored whether the pattern of temporary mobility was a predictor. First, we tested an independence baseline model to predict usual residence (U). The predictors consisted of a mobility variable (M), sex (S), and the amount of time spent away (T). The mobility variable was comprised of the four mobile categories (1-Shots, No-Repeats, Boomerangs, and Floaters). Amount of time spent away was split into two categories: less than half the reference period and half or more of the reference period. Race and age were excluded since neither improved the fit of the data.

Table 5

Goodness-of-Fit Tests and Parameter Estimates for Log-Linear Models of the Effect of Sex (S), Temporary Mobility (M), and Length of Time Away (T) on Determination of Usual Residence Status (U)

A. Goodness of Fit Test

		(U) Usual Residence Status	
Model	*d.f.*	Chi-square +	*p*
1. U, SMT	15	4.79	.00
2. US, UM, UT, SMT	10	1.06	.12
3. UTM, USM, SMT	4	0.78	.16

B. Parameter Estimates, Model 3

	beta	s.e.	std. value
(M) MOBILITY:			
1-Shots	1.08	.40	2.71*
Boomerangs	-1.54	.39	-3.94*
No-Repeats	.83	.58	1.43
Floaters	-.38	.47	-.80
(S) SEX:			
(Males)	.39	.27	1.44
(T) TIME AWAY:			
(> ½ ref. period)	-1.78	.27	-6.52*
(U)*(S)*(M) INTERACTION (Males)			
1-Shots	-.64	.43	-1.48
Boomerangs	.69	.58	1.18
No-Repeats	.85	.62	1.37
Floaters	-.90	.42	-2.14*
(U)*(M)*(T) INTERACTION (> ½ ref. period)			
1-Shots	-.72	.48	-1.50
Boomerangs	-1.20	.54	-2.26*
No-Repeats	1.57	.74	2.12*
Floaters	.36	.41	0.88

+ Jackknife Pearson chi-square for overall fit.

* Significant at the .05 level.

The baseline model (U, SMT) did not fit the data well so we rejected the null hypothesis that assignment of usual residence is independent of mobility pattern, sex, and amount of time spent away (Jackknife X^2 overall goodness of fit = 4.79, *d.f.* = 15, *p* = .00, see Table 5). We then fitted a main effects model (2) which includes the additive effects of S, M and T upon U (US, UM, UT, SMT). This model yielded a good fit (Jackknife X^2 overall goodness of fit = 1.06, *d.f.* = 10, *p.* = .12). Lastly, a model (3) including two interaction terms was also fitted (UTM, USM, SMT). This model assumes interactive effects of T*M and of S*M on U. A comparison between the main effects and interaction model suggested that several interactions were significant and should be retained (comparison Jackknife X^2 = 1.99, *d.f.* = 6, *p* = .02). Table 5 contains the overall goodness of fit tests along with the parameter estimates from the best fitting interaction model (UTM, USM, SMT – Jackknife X^2 overall goodness of fit = 0.78, *d.f.* = 4, *p* = .16.)

The parameter estimates from Table 5 illustrate that temporary mobility has a significant main effect on assignment of usual residence in model 3 which controls for sex, amount of time spent away, and several interactions. Two of the mobility categories had significant beta coefficients albeit the directions were opposite. The 1-Shots were significantly more likely to be defined as usual residents (b = +1.08). Conversely, the Boomerangs had a negative parameter estimate (b = −1.54) meaning that the odds of being defined a usual resident were significantly decreased for this group.

Time spent away from the interview household had by far the largest effect on predicting usual residence with a strong negative association (b = −1.78). This means that for our temporarily mobile population, those away half or more of the reference period were significantly less likely to be considered usual residents than those away less than half of the time. Sex did not have a significant main effect, but was involved in a significant interaction. The interaction appears in the Floater group where male Floaters were less likely to be categorized as usual residents than female Floaters (b = −.90). Further investigation revealed few clues to explain this finding. Male and female Floaters differed little in the types of places they visited, their reasons for visiting, and the relation to the householder of places they visited (relative versus non-relative). Perhaps the interaction reflects differences in other social attachments such as presence of children, personal belongings, and/or contribution of resources.

The bottom of table 5 indicates that the interaction between usual residence, mobility and amount of time spent away is rather complex. The amount of time spent away appears to affect usual residence status for some types of mobility but not for others. The interaction coefficient is significant and negative for the Boomerangs (b = −1.20). Thus, the odds of being defined a usual resident are even lower for Boomerangs away half or more of the reference period compared to other groups away for a similar amount of time. This suggests that persons who "boomerang" back and forth between two households will be considered usual residents at the place they spend the majority of time.

However, for the No-repeats, the coefficient is significant and *positive*, essentially canceling out time away's negative main effect (1.57 + −1.78 = −0.21). For this group, the amount of time spent away appears to have no association with usual residence assignment. Apparently, factors other than time may be more important in the cognitive process of determining where these persons "reside." One hypothesis is that No-repeaters are persons who must travel for a living and who, despite their frequent mobility and long periods away, clearly "belong" to a stable residence. This notion supports findings from a vignette study that found respondents did not require a stated rule to be able to correctly identify the usual residence of persons described as being away on business travel. Such persons were "intuitively" perceived to be part of the households from which they were away (Gerber, Wellens and Keeley 1996).

7. CONCLUSIONS

Temporary mobility, as defined in our research, involves long and short, frequent and infrequent, patterned and unpatterned movement away from, but often back to, a single residence. Such mobility has long been hypothesized to contribute toward census and survey coverage error by blurring the concept of who exactly lives or stays at a particular household.

Our sample of persons having a more-than-casual association to households indicated a fair amount of temporary mobility over a two-three month period. Interesting demographic differences were noted in the level of mobility as well as the pattern of mobility reported. The "hard to enumerate" (HTE) group (black/Hispanic males between 18 and 29) were found to cluster in the Boomerang and Floater groups, suggesting a repeat pattern of temporary mobility. We suspect these groups include persons having strong attachments to multiple households, for example an adult son who splits time between a parent and girlfriend's or a young mother who stays periodically at different kin-network households to receive assistance with child care.

Besides the inclusion of the types of persons who tend to be missed in censuses and surveys, other considerations point to the Boomerang and Floaters as being of particular interest. First, compared to the other mobility categories, these groups spent a longer time away from the households in which they were "found" and second, were more often classified as having a usual residence someplace other than the household in which they were found. It is difficult to estimate how much this type of mobility contributes toward undercounting. However, it is very noteworthy that half the HTE population fall in either the Boomerang or Floater group. It seems more than a coincidence that such a large segment of this population belong to one of the two mobility groups most easily labeled "residentially ambiguous."

The log-linear analysis suggests that there is not a clearcut, simple relationship between temporary mobility and assignment of usual residence. We do not find that the greater the amount of temporary mobility the less the chance of being defined a usual resident. Instead, the relationship seems more driven by the pattern of movement. For example, the traveling salesman or truck driver who reports the greatest variety of places visited and the largest number of visits may, nonetheless, have less residential ambiguity than a person visiting only one other place but making many repeat visits. And, in fact, this proved to be the case for the No-Repeats for whom the amount of time spent away had essentially no relation to usual residence assignment.

Our exploration of temporary mobility represents a new research direction for the study of within-household census and survey coverage error. Two recommendations for improving census and survey coverage are offered. First, survey organizations should explore the possibility of directly measuring the association between temporary mobility and incidents of census and survey undercoverage. This could be accomplished by adding questions about mobility to post-census coverage interviews used to estimate the number of people missed or counted in error. If the correlation between coverage error and mobility is significant, then survey methods and procedures could be adjusted to try and reduce it. For example, new roster probes could be added to census forms and nonresponse follow-up interviews, the aim being to find more of the Boomerangs and Floaters. Measures of temporary mobility might also prove to be a powerful predictor variable when statistically modeling the undercount. While admittedly in the early stages, temporary mobility looks promising as an avenue to better understanding household coverage error.

ACKNOWLEDGMENTS

The authors wish to thank Elizabeth A. Martin, Theresa J. DeMaio, Robert E. Fay and three anonymous reviewers for insightful comments on earlier versions of this paper. 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

CHAKRABARTY, R. (1992). Coverage of the Current Population Survey (CPS) Relative to the 1990 Census. Unpublished U.S. Bureau of the Census memorandum to the record, February 20, 1992.

DE LA PUENTE, M. (1993). Why are people missed or erroneously included by the census: a summary of findings from ethnographic coverage reports. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 29-66.

DRESSLER, W., HOEPPNER, S., and PITTS, B. (1985). Household structure in a southern black community. *American Anthropologist*, 87, 835-862.

ELLIS, Y. (1994). Categorical Data Analysis of Census Omissions, Internal Memorandum, Washington D.C: U.S. Bureau of the Census.

ELLIS, Y. (1993). Census Error Study. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 248.

FAY, R.E. (1989a). An analysis of within-household undercoverage in the current population survey. *Proceedings of the 1989 Annual Research Conference*, U.S. Bureau of the Census, 156-175.

FAY, R.E. (1989b). CPLX: Contingency Table Analysis for Complex Sample Designs. Program Documentation. Unpublished document, U.S. Bureau of the Census.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FEIN, D.J., and WEST, K. (1988). Toward a theory of coverage error: an exploration of data from the 1986 Los Angeles test census. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 540-562.

GERBER, E., WELLENS, T., and KEELEY, C. (1996). Who Lives Here?: The Use of Vignettes in Household Roster Research. Paper presented at the annual meeting of the American Association for Public Opinion Research, Salt Lake City.

GERBER, E. (1994). The Language of Residence: Respondent Understandings and Census Rules. Unpublished report of the Cognitive Study of Living Situations. Center for Survey Methods Research, U.S. Bureau of the Census.

GRIFFIN, D., and MORIARITY, C. (1992). Characteristics of Census Errors. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 179.

HAINER, P. (1987). A Brief and Qualititative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households. Report for the Census for Survey Methods Research, U.S. Bureau of the Census, April 8, 1987.

HAINER, P., HINES, C., MARTIN, E.A., and SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 513-539.

HANSEN, K. (1994). Geographical Mobility: March 1993 to March 1994. Current Population Reports, Population Characteristics P20-485. U.S. Department of Commerce.

HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.

HUDGINS, J.L., and HOLMES, B.J. (1993). The impact of social and economic marginality on the underenumeration of African American males. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 153-166.

LIPTON, S.G., and ESTRADA, L.F. (1993). Factors associated with undercount rates in Los Angeles county. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 83-102.

LYNCH, J.T., WITT, M., BRANSON, S., and ARDINI, M. (1993). Living Situation Survey: Final Methods Report, Unpublished report, Research Triangle Institute: Research Triangle Park.

MAHLER, S. (1993). Alternative Enumeration of Undocumented Salvadorans on Long Island. Prepared under Joint Statistical Agreement 89-46 with Columbia University. U.S. Bureau of the Census, Washington, D.C.

MARTIN, J.M., and MARTIN, E.P. (1985). The Helping Tradition in the Black Family and Community. Silver Spring, Md.: National Association of Social Workers.

MORIARITY, C.L., and CHILDERS, D. (1993). Analysis of Census Omissions: Preliminary Results. U.S. Bureau of the Census, DSSD 1990 REX Memorandum Series #PP-8.

MONTOYA, D. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount: Woodburn, Oregon. Ethnographic Evaluation of the 1990 Decennial Census Report #10. Prepared under Joint Statistical Agreement 89-30 with the University of Oklahoma. U.S. Bureau of the Census: Washington, D.C.

PENNIE, K. (1990). Coverage Comparisons Between the 1990 Census and Current Population Survey (CPS). Unpublished U.S. Bureau of the Census memorandum to Preston Jay Waite.

ROBINSON, J.G, AHMED,B., DAS GUPTA, P., and WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1071.

ROMERO, M. (1992). Ethnographic Evaluation of the Behavioral Causes of Census Undercount of Undocumented Immigrants and Salvadorans in the Mission District of San Francisco, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #18. Prepared under Joint Statistical Agreement 89-41 with the San Francisco State University Foundation. U.S. Bureau of the Census, Washington, D.C.

SHAPIRO, G., DIFFENDAL, G., and CANTOR, D. (1993). Survey undercoverage: major causes and new estimates of magnitude. *Proceedings of the 1993 Annual Research Conference*, U.S. Bureau of the Census, 638-663.

STACK, C.B. (1974). *All Our Kin: Strategies for Survival in the Black Community.* New York: Harper and Row.

SWEET, E.M. (1994). Roster research results from the living situation survey. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 415-433.

U.S. DEPARTMENT OF LABOR (1994). Employment and Earnings. Bureau of Labor Statistics, January 1994: Washington, D.C.

VALENTINE, C., and VALENTINE, B. (1971). Missing Men: A Comparative Methodology Study of Underenumeration and Related Problems. Report to the U.S. Bureau of the Census, May 3, 1971.

VELASCO, A. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount in the Community of Sherman Heights, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #22. Prepared under Joint Statistical Agreement 89-42 with the Chicano Federation of San Diego County. U.S. Bureau of the Census, Washington, D.C.

WINGERD, J. (1992). Urban Haitians: Documented/Undocumented in a Mixed Neighborhood. Ethnographic Evaluation of the 1990 Decennial Census, Report #7. Prepared under Joint Statistical Agreement # 90-10 with the Community Service Council of Broward County, Inc. U.S. Bureau of the Census, Washington, D.C.

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 14, Number 1, 1998

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.

CONTENTS TABLE DES MATIÈRES

Volume 26, No. 1, March/mars 1998

CONTENTS TABLE DES MATIÈRES

Volume 26, No. 2, June/juin 1998

First Announcement

Call for Papers

IASS SATELLITE CONFERENCE ON SMALL AREA ESTIMATION Riga, Latvia, 20-21 August 1999

The Satellite Conference on Small Area Estimation will follow the ISI session in Helsinki. It is intended to cover aspects of both theoretical background in small area estimation and practical application of different estimation methods for small area statistics. This includes sample design for small area statistics (national experiences), new developments in the field of estimation for small area statistics and successful applications of small area estimation techniques, including those that use data from administrative systems. Small area statistics is a subject of great interest in many countries. Several statistical agencies in Western countries have introduced vigorous programmes to meet this new demand, with a view toward producing efficient and high quality statistics. Several international conferences and seminars have been organised in the last years and others are yet to be organised. Furthermore, significant research on both the theoretical and practical aspects of small area estimation is conducted at various universities and some national statistical offices.

The Conference is organised on the initiative of the Baltic countries, and is aimed at improving knowledge transfer of new methods. The proceedings of the Conference should be of interest to all statisticians working in this field but it is of particular interest for the economies in transition in Central and Eastern European countries and the former Soviet Union countries, where complete reporting and complete statistical investigations are to be replaced or have been replaced with sample surveys, the production of reliable small area statistics has emerged as a pressing and frequently difficult and costly problem.

The conference proceedings will be opened by Dr. Danny Pfeffermann, who will provide an overview of the New Developments in Small Area Estimation. Initial plans also include holding a one day Short Course on Small Area Estimation immediately preceding the Conference, in order to allow some participants to acquire the basic knowledge that would allow them to appreciate fully the proceedings of the conference. The meeting is sponsored by the International Association of Survey Statisticians (IASS), the Central Statistical Bureau of Latvia (CSBL), and the University of Latvia (UL).

The members of the International Programme Committee are: Ödon Éltetö (Hungary), Wayne A. Fuller (USA), Jan Kordos (Poland, chair), John Kovar (Canada), Juris Krumins (Latvia), Janis Lapinš (Latvia), Danny Pfeffermann (Israel), Richard Platek (Canada), J.N.K. RAO (Canada), Carl-Erik Särndal (Canada), Dennis Trewin (Australia) and Janusz Wywial (Poland).

Abstracts of proposed papers should include full information on authors and their affiliations, and the contact address (including e-mail and fax) and a text of 200-300 words. The deadline for submission is December 31, 1998. Earlier submissions are encouraged and notifications of acceptance will be sent out as soon as possible. Acceptance is conditional on the attendance of the meeting by at least one of the authors. Abstract should be submitted, preferably via e-mail (in ASCII or WORD 6.0), or by fax or by mail to:

Jan Kordos, Al. Niepodleglosci 208, 00-925 Warsaw, Poland;
Fax: (0048-22) 825-03-95; E-mail: kordos@gus.stat.gov.pl
or to any other members of the Programme Committee

It is the intention of the Programme Committee to publish the papers presented at the Conference in a special Proceedings of the Conference issue. The papers may also be published in any journal after the Conference.

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.