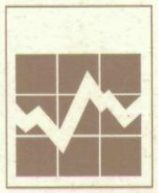

SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1994

•

VOLUME 20

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1994 • VOLUME 20 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

December 1994

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

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 20, Number 2, December 1994

CONTENTS

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 95–96
Statistics Canada

In This Issue

This issue of *Survey Methodology* opens with a special section on **Establishment Survey Methods**. The four papers in this special section deal with important issues in the context of establishment surveys such as questionnaire design, sample design and estimation. These papers were initially presented at the International Conference on Establishment Surveys, Buffalo, New York, June 1993.

The paper by Armstrong and St-Jean presents an application of the general framework of regression estimation in two-phase sampling. Using data from a two-phase sample of tax records, three particular cases of the generalized regression estimator - two regression estimators and a poststratified estimator - are empirically compared to the Horvitz-Thompson estimator. The empirical study shows that the poststratified estimator is more efficient than the Horvitz-Thompson estimator and as efficient as the two regression estimators.

Gallego, Delincé and Carfagna describe the Monitoring Agriculture with Remote Sensing (MARS) project of the European Community. As the project is not capable of producing good estimates of crop areas and yields, they describe a method of sampling farms by points to obtain reliable estimates. Results of applying this approach in two regions, Emilia Romagna in Italy and the Czech Republic, are described.

Pollock, Turner and Brown discuss the use of capture-recapture sampling to estimate the population size and population totals when only incomplete list frames exist. A discussion of the properties of the resulting model based estimators and an example where the establishments are fishing boats are presented.

In the last paper of this special section, Gower gives an overview of important considerations that should be taken into account when developing and designing questionnaires for business surveys. Examples of applications of focus groups and cognitive research to test questionnaires for business surveys are presented.

Rancourt, Lee and Särndal present simple correction factors to reduce the bias of the standard estimator of the population mean in the case of ratio imputation for confounded nonresponse. The effectiveness of these factors is studied by Monte Carlo simulations. The factors are found to be effective especially when the model underlying ratio imputation holds.

The use of the capture-recapture approach for coverage evaluation of the U.S. census is discussed by Ding and Fienberg. They give methods for estimating population total and census undercount when the assumption of a perfect match between individuals in the census and in the sample is relaxed. They propose models to describe two types of matching errors, mismatches and erroneous non-matches. The methods are illustrated using data from 1986 Los Angeles test census and 1990 Decennial Census.

Kott discusses testing a hypothesis about linear regression coefficients using data from a sample survey. He suggests an adjustment of the design-based linearization variance estimator to reduce its model bias and a formula to estimate its effective degrees of freedom. Two examples of the method are presented.

Cox develops a framework, called matrix masking, for microdata disclosure limitation methods that should improve understanding of these methods and of their effect on data use. Within this framework, based on ordinary matrix arithmetic, statistical agencies can develop, evaluate and use reliable software for disclosure limitation of microdata. The author presents explicit matrix mask formulations for the principal microdata masking methods in current use.

Falorsi, Falorsi and Russo conduct an empirical comparison of some small area estimation methods in the context of the Italian Labour Force Survey using data from the 1981 Italian Census. The estimators included in their study are a poststratified direct estimator, a synthetic estimator, an optimal linear combination of the two, and a sample size dependent estimator. They conclude that, for their application, the sample size dependent estimator offers the best balance of variance and bias.

The paper by Niyonsenga presents a comparison of two nonparametric methods of estimation of response probabilities in sampling theory via a Monte Carlo simulation. It is shown that, in the context of simple random sampling without replacement, the nonparametric variant based on the ranks of the values of the auxiliary variable performs better, with respect to both bias and mean square error, than the method based on the values of the auxiliary variable, for both the expansion and regression estimators.

Schabenberger and Gregoire compare alternative exact and approximate πps strategies in the context of sampling in forestry. Two sequential sampling schemes due to Sunter combined with the Horvitz-Thompson estimator are compared to the random group strategy of Rao, Hartley and Cochran (RHC) as well as a ratio of means estimator used with simple random sampling. If the size variable is highly correlated with the variable of interest then πps strategies are considerably more efficient. When the correlation is very high the exact πps strategy is most efficient; however, the RHC strategy has the advantage of simplicity. If the correlation is low then the πps strategies can be very inefficient.

The Editor

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 97–105
Statistics Canada

Generalized Regression Estimation for a Two-Phase Sample of Tax Records

JOHN ARMSTRONG and HÉLÈNE ST-JEAN[1]

ABSTRACT

A generalized regression estimator for domains and an approximate estimator of its variance are derived under two-phase sampling for stratification with Poisson selection at each phase. The derivations represent an application of the general framework for regression estimation for two-phase sampling developed by Särndal and Swensson (1987) and Särndal, Swensson and Wretman (1992). The empirical efficiency of the generalized regression estimator is examined using data from Statistics Canada's annual two-phase sample of tax records. Three particular cases of the generalized regression estimator - two regression estimators and a poststratified estimator - are compared to the Horvitz-Thompson estimator.

KEY WORDS: Model assisted estimation; Domain estimation; Poisson sampling.

1. INTRODUCTION

In this paper the problem of domain estimation under two-phase sampling for stratification is examined in a case in which Poisson sampling is used at both phases of selection. Consider a population of N units and suppose that it is necessary to estimate the total of a characteristic of interest, y, for L disjoint domains. Domain membership can be well, but not exactly, predicted using an auxiliary variable, Θ, that is not observed before sampling. The cost of obtaining information on Θ is lower than the cost of obtaining information on y and lower than the cost of obtaining exact domain membership data. At the first phase of sampling, a Poisson sample is drawn from the population and the value of Θ is observed for each sampled unit. The units in the first-phase sample are stratified using Θ-values. This stratification is an approximation to stratification by domain. At the second phase of sampling, a Poisson sample is drawn from each stratum. The value of y, as well as exact domain membership data, is observed for each unit in the second-phase sample.

The Horvitz-Thompson estimator of the total of y for domain d is $\hat{Y}_{H-T}(d) = \sum_{i \epsilon s2} y_i(d)/(p_{1i}p_{2i})$, where $y_i(d)$ takes the value of y_i if unit i falls in domain d and otherwise takes the value zero, $s2$ denotes the second-phase sample and p_{1i} and p_{2i} are first- and second-phase selection probabilities, respectively, for unit i. Since the sample sizes obtained using Poisson sampling are random variables, this estimator may be inefficient. (See Sunter 1986 or Särndal, Swensson and Wretman 1992, p. 63.) Generalized regression estimation is an alternative to the Horvitz-Thompson estimator that can be employed when auxiliary information is available. A generalized regression estimator for two-phase Poisson sampling and an approximate estimator of its variance are derived in this paper.

Section 2 contains the derivation of the generalized regression estimator and approximate variance estimator. Section 3 includes a description of the application that motivated the estimation problem - Statistics Canada's annual two-phase sample of tax records. The results of an empirical study comparing the Horvitz-Thompson estimator with three particular cases of the generalized regression estimator - the poststratified estimator currently used in production and two regression estimators - are described in Section 4.

2. GENERALIZED REGRESSION ESTIMATION

Generalized regression estimation is not a new technique. A generalized regression estimator for a one-phase sample design is described by Deming and Stephan (1940). Recent applications of generalized regression estimation at Statistics Canada include the work of Lemaître and Dufour (1987) and Bankier, Rathwell and Majkowski (1992). Hidiroglou, Särndal and Binder (1993) provide an extensive discussion of the use of generalized regression estimators for business surveys.

Derivation of generalized regression estimators can be approached from the perspective of model assisted survey sampling (Särndal, Swensson and Wretman 1992) or from the perspective of calibration (Deville and Särndal 1992). Let $U = \{u\}$ and $V = \{v\}$ denote sets of first-phase poststrata and second-phase poststrata, respectively. During generalized regression weighting of the first-phase sample, the design weights $1/p_{1i}$ are adjusted to yield weights $w_{1i} = g_{1i}/p_{1i}$ that respect the calibration equations

[1] John Armstrong, Social and Economic Studies Division, 24 - R.H. Coats Bldg., and Hélène St-Jean, Business Survey Methods Division, 11 - R.H. Coats Bldg., Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

$$\sum_{i \in s1 \cap v} w_{1i} x_i = X_u,$$

for each first-phase poststratum u, where x_i is an $L_1 \times 1$ vector of auxiliary variables known for all units in the population and X_u is the vector of auxiliary variable totals for poststratum u. The adjusted weights minimize the distance measure $\sum_{i \in s1} (g_{1i} - 1)^2 / p_{1i}$. The same weights can be obtained from a model assisted perspective using

$$E_\xi(y_i) = x_i' \beta_u, \; i \in u$$
$$V_\xi(y_i) = \sigma^2,$$

where y_i is the value of the variable of interest for unit i, and $E_\xi(\cdot)$ and $V_\xi(\cdot)$ denote expectation and variance, respectively, with respect to the model.

For the generalized regression estimators of interest, weighting of the second-phase sample involves a calibration procedure that is conditional on the results of first-phase weighting. The initial weights, w_{1i}/p_{2i}, are adjusted to give final weights, $w_i = g_{2i} w_{1i}/p_{2i}$, that satisfy the calibration equations

$$\sum_{i \in s2 \cap v} w_i z_i = \tilde{Z}_v,$$

for each second-phase poststratum v, where z_i is an $L_2 \times 1$ vector of auxiliary variables known for all units in the first-phase sample and $\tilde{Z}_v = \sum_{i \in s1 \cap v} w_{1i} z_i$ is an estimate of the vector of auxiliary variable totals for post-stratum v, computed using the adjusted first-phase weights w_{1i}. Note that these calibration equations differ in an important way from the examples given by Särndal and Swensson (1987, pp. 284-288) and Särndal, Swensson and Wretman (1992, pp. 359-366) because they involve adjusted first-phase weights rather than first-phase design weights. The final weights minimize the distance measure $\sum_{i \in s2} w_{1i}(g_{2i} - 1)^2 / p_{2i}$. The model needed to obtain the same weights from a model assisted perspective is

$$E_\xi(w_{1i} y_i) = w_{1i} z_i' \beta_v, \; i \in v$$
$$V_\xi(w_{1i} y_i) = w_{1i} \sigma^2.$$

Use of adjusted first-phase weights rather than first-phase design weights in the second-phase calibration equations has two important advantages. First, the generalized regression estimator for domain d can be written as

$$\hat{Y}_{\text{GREG}}(d) = \sum_{i \in s2} y_i(d) g_{1i} g_{2i} / p_{1i} p_{2i},$$

using first-phase and second-phase g-weights. Second, suppose that some auxiliary variables are used for calibration at both phases of weighting. Estimates of population totals for such variables that are equal to actual totals can be constructed using final weights.

Let $\check{X}_u = \sum_{i \in s1 \cap u} x_i / p_{1i}$ denote the $L_1 \times 1$ vector of Horvitz-Thompson estimates of auxiliary variable totals for first-phase poststratum u. The first-phase g-weight is

$$g_{1i} = 1 + \lambda_u' x_i,$$

where $\lambda_u' = (X_u - \check{X}_u)' M_u^{-1}$ and $M_u^{-1} = (\sum_{i \in s1 \cap u} x_i x_i' / p_{1i})^{-1}$. For second-phase poststratum v, denote the estimate of $\tilde{Z}_v$ based on initial second-phase weights by $\dot{Z}_v = \sum_{i \in s2 \cap v} w_{1i} z_i / p_{2i}$. The second-phase g-weight is

$$g_{2i} = 1 + \lambda_v' z_i,$$

where $\lambda_v' = (\tilde{Z}_v - \dot{Z}_v)' M_v^{-1}$ and $M_v^{-1} = (\sum_{i \in s2 \cap v} w_{1i} z_i z_i' / p_{2i})^{-1}$.

The approximate variance of $\hat{Y}_{\text{GREG}}(d)$ is given by

$$V(\hat{Y}_{\text{GREG}}(d)) \approx \sum_i \frac{1 - p_{1i}}{p_{1i}} Q_{1i}^2 + E_1\left[\sum_{i \in s2} \frac{1 - p_{2i}}{p_{2i}} (w_{1i} Q_{2i})^2 \right],$$

where $E_1(\cdot)$ denotes expectation with respect to the first phase of sampling, $Q_{1i} = y_i(d) - x_i' B_u$ for each unit in first-phase poststratum u, and B_u, the vector of estimated coefficients from the regression of $y(d)$ on x that would be obtained if $y(d)$ was available for all units in first-phase poststratum u, is given by

$$B_u = \left(\sum_{i \in u} x_i x_i' \right)^{-1} \left(\sum_{i \in u} x_i y_i(d) \right).$$

Similarly, $Q_{2i} = y_i(d) - z_i' B_v$ for each unit in second-phase poststratum v and B_v, the vector of estimated coefficients from the regression of $y(d)$ on z that would be obtained, conditional on the first-phase calibration, if $y(d)$ was available for all units in the component of the first-phase sample falling in second-phase poststratum v, is given by

$$B_v = \left(\sum_{i \in s1 \cap v} w_{1i} z_i z_i' \right)^{-1} \left(\sum_{i \in s1 \cap v} w_{1i} z_i y_i(d) \right).$$

An estimator of the approximate variance of $\hat{Y}_{\text{GREG}}(d)$ is

$$\hat{V}(\hat{Y}_{\text{GREG}}(d)) = \sum_i \frac{1 - p_{1i}}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2 + \sum_i \frac{1 - p_{2i}}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Since $y(d)$ is available only for units in $s2$, estimates of B_u and B_v are

$$\hat{B}_u = \left(\sum_{i \in s2 \cap u} w_i x_i x_i' \right)^{-1} \left(\sum_{i \in s2 \cap u} w_i x_i y_i(d) \right),$$

$$\hat{B}_v = \left(\sum_{i \in s2 \cap v} w_i z_i z_i' \right)^{-1} \left(\sum_{i \in s2 \cap v} w_i z_i y_i(d) \right).$$

The sample residuals needed to compute the variance estimator are $q_{1i} = y_i(d) - x_i' \hat{B}_u$ and $q_{2i} = y_i(d) - z_i' \hat{B}_v$. More details of the derivation of the approximate variance of $\hat{Y}_{\text{GREG}}(d)$ and the estimator of the approximate variance are given in Appendix A.

If y is strongly correlated with x and z, the variance of the generalized regression estimator of the population total of y will be relatively small. However, it is important to note that strong correlations between y and x and z will not necessarily lead to a relatively small variance for the estimate of the total of y for a particular domain, since $y(d)$ may be poorly correlated with x and z within poststrata that include at least one sampled unit falling in domain d.

The correlation between $y(d)$ and x and z within a poststratum that includes at least one sampled unit falling in domain d may be low if some sampled units in the poststratum do not fall in domain d. This situation may arise often if domain totals of auxiliary variables and/or exact domain membership information for units in the first-phase sample are unavailable. In the context of two-phase sampling for stratification, there is no domain membership information available before selection of the first-phase sample. If each first-phase poststratum is formed by combining one or more first-phase sampling strata, for example, most first-phase poststrata will include more than one domain. The variable Θ used to predict domain membership during stratification of the first-phase sample is not an exact predictor. If second-phase poststrata are formed by combining second-phase sampling strata, each domain may be divided between a number of second-phase poststrata.

Depending on the type of auxiliary information used, the g-weights associated with the generalized regression estimator and, consequently, generalized regression estimates, may be negative.

3. APPLICATION: TWO-PHASE SAMPLING OF TAX RECORDS

The two-phase tax sample is part of a general strategy at Statistics Canada for production of annual estimates of Canadian economic activity. Annual economic data for large businesses are collected through mail-out sample surveys. Data for small businesses are obtained from the tax sample. Estimates of financial variables for the business population are obtained by combining tax and survey estimates. Tax data rather than survey data are used to obtain small business estimates in order to reduce costs and response burden.

The two-phase sample design was introduced in response to a requirement for estimates for domains defined using the four-digit Standard Industrial Classification (SIC) code (Statistics Canada 1980). The first two digits of SIC (SIC2) provides a classification of businesses activity into 76 groups. Within each group, four-digit SIC (SIC4) codes provide classification into finer categories. For example, the SIC2 code of a business might classify it in the transportation industry while the SIC4 code describes the activity of the business as bulk liquids trucking.

There are two types of taxfilers - T1s and T2s. A T1 taxfiler is an individual, who may own all or part of one or more unincorporated businesses, while a T2 taxfiler is an incorporated business. Administrative files that contain limited information for all taxfilers that are associated with businesses are provided to Statistics Canada by Revenue Canada, the Canadian government department responsible for tax collection. These files are used to construct a sampling frame. Information concerning numbers of businesses owned by T1 taxfilers and ownership shares is not available on the sampling frame. Frame data does include geographical information, as well as gross business income and net profit for both T1 and T2 taxfilers. A few other major financial variables, including salary and inventory data, are generally available for T2 taxfilers. Estimates are required for about 35 financial variables that can be obtained from tax returns and associated financial statements but are not available on administrative files supplied by Revenue Canada.

Taxfilers that are associated with businesses are classified by Revenue Canada using the SIC system. In most cases, descriptions of business activity reported on tax returns are sufficient to accurately determine SIC2 codes. Revenue Canada assigns additional digits of SIC to most taxfilers. However, not all taxfilers are classified to the four-digit level and the third and fourth digits of SIC4 codes assigned by Revenue Canada are relatively inaccurate. A two-phase approach to sampling of tax records was adopted to facilitate accurate estimation of economic production at the SIC4 level.

Section 3.1 includes a brief description of the two-phase sampling design. More information about the two-phase design is provided in Armstrong, Block and Srinath (1993). Sections 3.2 and 3.3 contain information concerning estimation for the two-phase design. The Horvitz-Thompson estimator is described in Section 3.2 and a poststratified estimator is discussed in Section 3.3.

3.1 Sampling Design

The administrative information used to construct the sampling frame for a particular tax year is accumulated by Revenue Canada over a period of two calendar years as tax returns are received and processed. The use of Poisson sampling offers substantial operational advantages because sampling operations can begin before a complete sampling frame is available.

The target (in-scope) population for tax sampling is the population of businesses with gross income over $25,000, excluding large businesses covered by mail-out sample surveys. The first-phase sample is a longitudinal sample of taxfilers. Strata are defined by SIC2, province and size (gross business income). All taxfilers that are included in the first-phase sample for tax year T and are still in-scope for tax sampling in tax year $T + 1$ remain in the first-phase sample for tax year $T + 1$. Taxfilers may be added to the first-phase sample each year to improve the precision of certain estimates and to replace taxfilers sampled in previous years that are no longer in-scope.

To implement Poisson sampling for first-phase sample selection, each taxfiler is assigned a pseudo-random number (hash number) in the interval (0,1) generated by a hashing function that uses the unique taxfiler identifier as input. The hash number for each taxfiler is compared to the sampling interval for the corresponding stratum. If the hash number for a particular taxfiler falls in the corresponding sampling interval and the taxfiler is not already in the first-phase sample, then the taxfiler is added to the first-phase sample. Since taxfiler identifiers do not change over time, Poisson sampling facilitates selection of a longitudinal first-phase sample.

First-phase selection probabilities for taxfilers that are already included in the first-phase sample are updated each year. Longitudinal updating is necessary because: (i) a taxfiler may fall in different first-phase sampling strata in consecutive tax years; and (ii) first-phase sampling fractions for a given stratum may vary from one year to the next.

Copies of tax returns and associated financial statements for taxfilers in the first-phase sample are sent to Statistics Canada from Revenue Canada. In order to select the second-phase sample, statistical entities are created using information about businesses corresponding to taxfilers in the first-phase sample. Let $J = \{j\}$ denote the population of businesses that is the target population for tax sampling. A statistical entity, denoted by (i,j), is created for every taxfiler-business combination in the first-phase sample. For each T1 taxfiler in the first-phase sample, data for all businesses wholly or partially owned by the taxfiler (including ownership shares) that are needed to create statistical entities are available from tax returns and associated financial statements. Since there is a one-to-one correspondence between businesses and T2 taxfilers, a single statistical entity is created for each T2 taxfiler in the first-phase sample.

For each tax year, statistical entities that have not appeared in previous tax samples are assigned SIC4 codes by Statistics Canada. These codes are determined using information supplementary to business activity descriptions reported on tax returns and are more accurate in digits three and four than codes assigned by Revenue Canada. For statistical entities that have appeared in previous tax samples, the SIC4 assigned earlier is carried forward.

Conceptually, the second-phase sample is a sample of businesses. Operationally, it is a sample of taxfilers selected using statistical entities. Statistical entities are stratified using SIC4 codes assigned by Statistics Canada, as well as province and size. The total revenue of business j is used as the size variable for statistical entity (i,j). If one statistical entity corresponding to a T1 taxfiler is selected for the second-phase sample, then all statistical entities corresponding to the taxfiler are selected. Consequently, the second-phase selection probability for statistical entity (i,j) depends only on i.

Second-phase sample selection is done by the Poisson sampling method using hash numbers generated from taxfiler identifiers. The hashing function used for second-phase sample selection is independent of the first-phase hashing function.

Data for about 35 financial variables are transcribed from tax returns and associated financial statements for taxfilers selected in the second-phase sample. SIC4 codes assigned by Statistics Canada are updated, if necessary, to ensure that all SIC4 codes used during tabulation of estimates correspond to the current tax year.

3.2 Horvitz-Thompson Estimator

The second-phase sample is a sample of businesses selected using statistical entities. Since some businesses are partnerships, more than one statistical entity may correspond to the same business. To construct estimates for the population of businesses, an adjustment for the effects of partnerships is required. If business j is a partnership, it will be included in the second-phase sample if any of the corresponding taxfilers are selected. The usual Horvitz-Thompson estimator must be adjusted for partnerships to avoid over-estimation. Let δ_{ij} denote the proportion of business j owned by taxfiler i and suppose that statistical entity (i,j) is selected for the second-phase sample. The data for business j is adjusted by multiplying it by δ_{ij} so that only the component of income and expense items corresponding to taxfiler i is included in estimates. Rao (1968a) describes a similar adjustment in a slightly different context.

Let y_j denote the value of the variable y for business j. The Horvitz-Thompson estimate of the total of y over domain d, incorporating adjustment for partnerships, is given by

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} y_j(d) / (p_{1i} p_{2i}),$$

where J_i is a set containing the indices of the businesses wholly or partially owned by taxfiler i. Since selection probabilities depend only on the taxfiler index i, $\hat{Y}_{H-T}(d)$ can be written as

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d) / (p_{1i} p_{2i}),$$

where

$$y_i(d) = \sum_{j \in J_i} \delta_{ij} y_j(d).$$

$\hat{Y}_{H-T}(d)$ is an unbiased estimator of the population total of y for businesses in domain d. Refer to Rao (1968a).

The second-phase sample is obtained by Poisson subsampling of the first-phase Poisson sample. Consequently, the second-phase sample is also a Poisson sample and the variance of $\hat{Y}_{H-T}(d)$ is

$$V(\hat{Y}_{H-T}(d)) = \sum_i \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})\right] y_i(d)^2.$$

An unbaised estimator of this variance is

$$\hat{V}(\hat{Y}_{H-T}(d)) = \sum_{i \in s2} \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})^2\right] y_i(d)^2.$$

3.3 Poststratified Horvitz-Thompson Estimator

Adjustment of the Horvitz-Thompson estimator to account for differences between actual and expected sample sizes under Poisson sampling was suggested by Brewer, Early and Joyce (1972). The methodology currently used to produce estimates based on the two-phase tax sample incorporates such adjustments.

Ratio adjustments are applied within poststrata during weighting of both the first- and second-phase samples. Choudhry, Lavallée and Hidiroglou (1989) provide a general discussion of weighting for a two-phase Poisson sample using poststratified ratio adjustments. Suppose that first-phase poststratum u contains N_u taxfilers. An estimate of the number of taxfilers in the population that fall in first-phase poststratum u, based on the first-phase sample, is

$$\check{N}_u = \sum_{i \in s1 \cap u} (1/p_{1i}).$$

The poststratified first-phase weight for taxfiler i, $i \in u$ is

$$w_{1i} = (1/p_{1i})(N_u/\check{N}_u).$$

An estimate of the number of taxfilers in second-phase poststratum v, based on the first-phase sample, is

$$\tilde{N}_v = \sum_{i \in s1 \cap v} w_{1i}.$$

An alternative estimate, using only units in the second-phase sample, is

$$\dot{N}_v = \sum_{i \in s2 \cap v} w_{1i}/p_{2i}.$$

The poststratified second-phase weight for statistical entity (i,j) in poststratum v is

$$w_{2i} = (1/p_{2i})(\tilde{N}_v/\dot{N}_v)$$

and the final weight is

$$w_i = w_{1i} w_{2i}.$$

The poststratified estimate of the total of y over domain d is

$$\hat{Y}(d) = \sum_{i \in s2} w_i y_i(d).$$

Choudhry, Lavallée and Hidiroglou (1989) note that the variance of $\hat{Y}(d)$ is approximately given by

$$V(\hat{Y}(d)) \approx \sum_u \sum_{i \in u} \frac{(1 - p_{1i})}{p_{1i}} \left(y_i(d) - \frac{Y_u(d)}{N_u}\right)^2 + \sum_v \sum_{i \in v} \frac{(1 - p_{2i})}{p_{1i} p_{2i}} \left(y_i(d) - \frac{Y_v(d)}{N_v}\right)^2,$$

where $Y_u(d)$ and $Y_v(d)$ are population totals for the variable y over the portions of the domain d belonging to poststrata u and v respectively.

This variance is estimated by

$$\hat{V}(\hat{Y}(d)) = \sum_u \sum_v \left(\frac{N_u}{\check{N}_u}\right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} \left(y_i(d) - \frac{\hat{Y}_u(d)}{\hat{N}_u}\right)^2 + \sum_u \sum_v \left(\frac{N_u}{\check{N}_u}\right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} \left(y_i(d) - \frac{\hat{Y}_v(d)}{\hat{N}_v}\right)^2,$$

where the estimates $\hat{N}_u$ and $\hat{N}_v$ are calculated using final weights.

The inclusion of the factor $(N_u/\tilde{N}_u)^2(\tilde{N}_v/\dot{N}_v)^2$ can be motivated by an improvement in the conditional properties of the estimator (Royall and Eberhardt 1975). A variance estimator for the ratio estimator for a one-phase sample design including an analogous adjustment factor has also been studied by Wu (1982). Empirical work reported by Wu and Deng (1983) indicates that the coverage properties of confidence intervals based on the normal approximation are improved using the adjustment factor.

$\hat{Y}(d)$ is a particular case of $\hat{Y}_{\text{GREG}}(d)$ that can be obtained if a single auxiliary variable with value one for all taxfilers is employed during both first-and second-phase weighting. In this case, we have $g_{1i} = N_u/\tilde{N}_u$ for all taxfilers in first-phase poststratum u and $g_{2i} = \tilde{N}_v/\dot{N}_v$ for all taxfilers in second-phase poststratum v. Note that negative g-weights are precluded by this choice of auxiliary variables. The variance estimator $\hat{V}(\hat{Y}(d))$ differs in a minor way from the estimator $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ for this particular case of $\hat{Y}_{\text{GREG}}(d)$. The second-phase g-weight appears in the leading term of $\hat{V}(\hat{Y}(d))$ but does not appear in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$.

4. EMPIRICAL STUDY

In order to compare the performance of $\hat{Y}_{H-T}(d)$, $\hat{Y}(d)$ and $\hat{Y}_{\text{GREG}}(d)$, an empirical study was conducted using data from the province of Quebec for tax year 1989. Since the estimator $\hat{Y}(d)$ is a special case of $\hat{Y}_{\text{GREG}}(d)$, it will be called $\hat{Y}_{\text{GREG-TPH}}(d)$ in subsequent discussion. (TPH is an abbreviation for two-phase Hájek.) Two other generalized regression estimators were considered. In both cases, x and z contains a variable with value one for all taxfilers. One generalized regression estimator involves calibration on taxfiler revenue during second-phase weighting. (Taxfiler revenue is included as a second auxiliary variable in z.) The second estimator involves calibration on taxfiler revenue at both phases of weighting. (Taxfiler revenue is included as a second auxiliary variable in both x and z.) Estimates of domain totals computed using these two estimators are denoted by $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$, respectively, in subsequent discussion.

Estimates were produced for two variables of interest – transcribed revenue and total expenses. There are some conceptual differences between transcribed revenue and taxfiler revenue. For example, capital gains and extraordinary items are included in taxfiler revenue in many industries while they are excluded from transcribed revenue. In addition, taxfiler revenue contains more data capture errors than transcribed revenue since it is not subject to the same level of quality control.

The population used for the study included about 140,000 T2 taxfilers who reported over $25,000 in revenue for tax year 1989. The first- and second-phase selection probabilities used during sampling for production for tax year 1989 were employed. The first-phase sample included approximately 31,000 taxfilers and there were about 23,000 businesses in the second-phase sample. The correlation between taxfiler revenue and transcribed revenue for businesses in the second-phase sample was 0.969, while the correlation between taxfiler revenue and total expenses was 0.960.

Large proportions of units in the first- and second-phase samples were selected with certainty. All units with first-phase selection probability one were excluded from first-phase weighting and the corresponding g-weights were set to one. Units with second-phase selection probability one were treated analogously during second-phase weighting. There were 9,884 units in the first-phase sample with first-phase selection probabilities different from one and 910 units in the second-phase sample with second-phase selection probabilities different from one. Each first-phase poststratum consisted of one or more of the first-phase sampling strata used during sampling for 1989 production. These strata were defined using five revenue classes. All the sampling strata included in any particular first-phase poststratum corresponded to the same revenue class. Each first-phase poststratum contained a minimum of twenty sampled units. The use of a minimum sample size was motivated by concerns about the bias in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ when the number of sampled units used for estimation of regression coefficients is very small (Rao 1968b). If a first-phase sampling stratum included fewer than twenty sampled units, it was combined with sampling strata for similar SIC2 codes and the same revenue class until a poststratum containing at least twenty sampled units was obtained. Application of this procedure led to 166 first-phase poststrata. Second-phase poststrata were formed analogously, combining sampling strata for similar SIC4 codes to obtain a minimum sample size of twenty for each poststrata. There were 30 second-phase poststrata.

First and second-phase weights for $\hat{Y}_{\text{GREG-TPH}}(d)$, $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$ were calculated using a modified version of the SAS macro CALMAR (Sautory 1991). The set of first-phase sampling weights calculated for the GREG-R1R2 estimator included twelve negative weights. There were no negative second-phase weights calculated for either GREG-R2 or GREG-R1R2. (Negative weights are not possible for the GREG-TPH estimator.) Estimates of transcribed revenue and total expenses were produced for 77 SIC2 domains, 256 SIC3 domains and 587 SIC4 domains using the three GREG estimators, as well as $\hat{Y}_{\text{H-T}}(d)$. Since GREG-R1R2 did not produce any negative estimates, no measures were taken to modify the negative weights associated with the estimator.

Results of comparisons of the GREG-TPH and H-T estimators are presented in Table 1 and Table 2. The mean gains and mean losses reported in the tables are averages of ratios of coefficients of variation. The GREG-TPH estimator performs better than the H-T estimator for the

Table 1

Comparison of GREG-TPH and H-T Estimators for Transcribed Revenue, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.768	20	1.113
SIC3	175	0.909	81	1.082
SIC4	359	0.945	228	1.079

Table 2

Comparison of GREG-TPH and H-T Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.773	20	1.100
SIC3	175	0.910	81	1.082
SIC4	355	0.945	232	1.079

majority of domains. The gains obtained using GREG-TPH are particularly large for SIC2 domains. At the SIC4 level, the estimated coefficient of variation (CV) for the GREG-TPH estimate of total expenses is lower than the estimated CV for the H-T estimate for 60.5% of domains. In cases in which the estimated CV for GREG-TPH is lower it is 5.5% smaller, on average, than the estimated CV for H-T. When the estimated CV for GREG-TPH is higher it is 7.9% larger than the estimated CV for H-T, on average. In addition to the information in Tables 1 and 2, there is another reason to prefer GREG-TPH to H-T. Each year, tax return information for some sampled taxfilers is not received by Statistics Canada or is unusable because it does not include the necessary financial statements. Assuming that such cases of nonresponse are ignorable, the GREG-TPH estimator provides an automatic nonresponse adjustment.

The results in Tables 1 and 2 indicate that the relative performance of the GREG-TPH and H-T estimators are very similar for both variables of interest. The results of the other comparisons of estimators done as part of this empirical study did not depend on the variable of interest in any important way. Consequently, only results for total expenses are reported in subsequent tables.

The GREG-TPH estimator is compared to GREG-R2 and GREG-R1R2 in Tables 3 and 4. Based on estimated coefficients of variation, GREG-R2 performs slightly better than GREG-TPH. Since a large proportion of units in the second-phase tax sample have second-phase selection probability one and both GREG-R2 and GREG-TPH use the same auxiliary variables during first-phase weighting, the marginal differences between GREG-R2 and GREG-TPH are not surprising. Estimated CVs for GREG-R1R2 are generally smaller than estimated CVs for GREG-TPH and the relative performance of GREG-R1R2 improves as domain size increases. Nevertheless, GREG-R1R2 is superior to GREG-TPH for only 64% of SIC4 domains, and the average increase in estimated CVs for those domains in which GREG-R1R2 did worse than GREG-TPH is larger than the average decrease in estimated CVs for domains in which GREG-R1R2 performed better.

Table 3

Comparison of GREG-R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R2		No Difference	Losses Using GREG-R2	
	Number	Mean	Number	Number	Mean
SIC2	38	0.993	26	13	1.001
SIC3	58	0.991	158	40	1.002
SIC4	88	0.988	439	60	1.009

Table 4

Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	51	0.867	26	1.170
SIC3	160	0.934	96	1.093
SIC4	377	0.954	210	1.074

The results in Tables 3 and 4 indicate that, although the GREG-R1R2 estimator shows some promise, it would be inappropriate to completely replace the GREG-TPH estimator currently used in production by GREG-R1R2. The improvements obtained using GREG-R1R2 are relatively marginal, given the strong correlation between taxfiler revenue and total expenses. Larger improvements could be obtained if: (i) SIC codes used for first-and second-phase stratification were always consistent with SIC codes used to determine the domain membership of sampled units; and (ii) formation of first-and second-phase poststrata did not require combination of sampling strata to obtain a minimum sample size in each poststratum.

The results reported in Table 5 were obtained after SIC codes assigned to taxfilers by Revenue Canada and SIC codes used for stratification of the second-phase sample were changed for sampled units, where necessary, to eliminate inconsistencies between these codes and those

Table 5
Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation, No Misclassification

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	66	0.778	11	1.057
SIC3	184	0.916	72	1.047
SIC4	402	0.944	185	1.034

used to determine domain membership. A comparison of Tables 4 and 5 indicates that the relative performance of GREG-R1R2 is considerably better when there are no classification errors. GREG-R1R2 reduces estimated CVs by over 22% (on average) for over 85% of SIC2 domains.

Throughout the empirical results reported here, performance improvements obtained through the use of additional auxiliary information increase as domain size increases. This result is consistent with the observations in Section 2 concerning the conditions under which correlations between $y(d)$ and the vectors of auxiliary variables, x and z, will be high. Provided that the variable of interest and the auxiliary variables are highly correlated, correlations involving $y(d)$ will be strong if each poststratum containing at least one sampled unit falling in domain d also contains relatively few sampled units that do not fall in domain d.

5. CONCLUSIONS

Generalized regression estimation provides a convenient framework for the use of auxiliary information. A generalized regression estimator for a two-phase sample design with Poisson sampling at both phases of selection is derived in this paper. The efficiency of the estimator is investigated through application to the two-phase tax sample selected by Statistics Canada to obtain annual estimates of the economic activity of small businesses. The estimation method currently used in production for this survey incorporates poststratified ratio adjustments during both first-and second-phase weighting to compensate for differences between actual and expected sample sizes. This poststratified estimator is a particular case of the generalized regression estimator.

In an empirical study, the generalized regression estimator currently used in production (GREG-TPH) performs much better then the Horvitz-Thompson estimator. Two other generalized regression estimators are also compared to GREG-TPH. The alternative estimators produce improvements for large domains. However, their performance for the smaller domains that are of particular interest to users of estimates based on the two-phase tax sample does not justify complete replacement of the current production methodology.

ACKNOWLEDGEMENTS

The authors would like to thank René Boyer for providing a modified version of the SAS macro CALMAR suitable for the empirical study, as well as K.P. Srinath and Michael Hidiroglou for helpful discussions. Thanks are also due to Michael Bankier and Jean Leduc for helpful comments on a earlier draft of this paper.

APPENDIX A: DERIVATION OF VARIANCE OF $\hat{Y}_{GREG}(d)$ AND VARIANCE ESTIMATOR

The variance of $\hat{Y}_{GREG}(d)$ can be derived using the identity

$$V(\hat{Y}_{GREG}(d)) = E_1V_2(\hat{Y}_{GREG}(d)) + V_1E_2(\hat{Y}_{GREG}(d)).$$

First, consider the variance of the estimator with respect to the second phase of sampling, conditional on the results of first-phase calibration. If the vector of auxiliary variables for second-phase weighting, z, includes a variable with value one for all taxfilers (or a linear combination of auxiliary variables that is equal to one for all taxfilers can be constructed), the generalized regression estimator can be written as

$$\hat{Y}_{GREG}(d) = \sum_{i\in s2} w_{1i}w_{2i}y_i(d)$$

$$= \sum_v \sum_{i\in s2\cap v} w_{1i}(y_i(d) - z_i'\hat{B}_v)/p_{2i} + \sum_v \hat{Z}_v\hat{B}_v.$$

Ignoring the variability due to the estimation of regression coefficients during second-phase weighting, we have

$$E_1V_2(\hat{Y}_{GREG}) \approx E_1V_2\left(\sum_{i\in s2} w_{1i}Q_{2i}/p_{2i}\right)$$

$$= E_1\left(\sum_{i\in s1} \frac{(1-p_{2i})}{p_{2i}} w_{1i}^2Q_{2i}^2\right).$$

The estimator of $E_1V_2(\hat{Y}_{GREG}(d))$ based on the variance estimator for calibration estimators advocated by Deville and Särndal (1992, p. 380) is

$$\hat{S}_1 = \sum_{i\in s2} \frac{(1-p_{2i})}{(p_{1i}p_{2i})^2} (g_{1i}g_{2i}q_{2i})^2.$$

Ignoring variability due to the estimation of regression coefficients during first-phase weighting, the second term in the variance expression can be written as

$$V_1 E_2(\hat{Y}_{\text{GREG}}(d)) = V_1\left(\sum_{i\in s1} w_{1i} y_i(d)\right)$$

$$= \sum_i \frac{(1 - p_{1i})}{p_{1i}} Q_{1i}^2.$$

An estimator of this term is

$$\hat{S}_2 = \sum_{i\in s2} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2.$$

REFERENCES

ARMSTRONG, J., BLOCK, C., and SRINATH, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.

BANKIER, M., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys.*

BREWER, K.R.W., EARLY, L.J., and JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.

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

HIDIROGLOU, M.A., SÄRNDAL, C.-E., and BINDER, D.A. (1993). Weighting and estimation in establishment surveys. Paper presented at the International Conference on Establishment Surveys, Buffalo, New York.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

RAO, J.N.K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

RAO, J.N.K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.

ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, Series C, 37, 43-52.

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

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

SAUTORY, O. (1991). La macro SAS: CALMAR. Unpublished manuscript, Institut national de la statistique et des études économiques, Paris.

STATISTICS CANADA (1980). *Standard Industrial Classification*. Catalogue No. 12-501E, Statistics Canada.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

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

WU, C.F.J., and DENG, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Box, G.E.P. *et al.* (Eds.), *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 245-277.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 107–115
Statistics Canada

Two-Stage Area Frame Sampling on Square Segments for Farm Surveys

F.J. GALLEGO, J. DELINCÉ and E. CARFAGNA[1]

ABSTRACT

In the MARS Project (Monitoring Agriculture with Remote Sensing) of the E.C. (European Community), area frames based on a square grid are used for area estimation through ground surveys and high resolution satellite images. These satellite images are useful, though expensive, for area estimation: their use for yield estimation is not yet operational. To fill this gap the sample elements (segments) of the area survey are used as well for sampling farms with a template of points overlaid on the segment. Most often we use a fixed number of points per segment. Farmers are asked to provide global data for the farm, and estimates are computed with a Horvitz-Thompson approach. Major problems include locating farmers and checking for misunderstanding of instructions. Good results are obtained for area and for production of the main crops. Area frames need to be complemented with list frames (multiple frames) to give reliable estimates for livestock.

KEY WORDS: Area frame; Point sampling; Segment sampling; Farm sampling.

1. INTRODUCTION

The main purpose of this paper is to present the method used to sample farms in an area frame by the MARS (Monitoring Agriculture with Remote Sensing) Project of the European Community (EC). Sampling farms is not a central activity in this project, but rather a way of bypassing the limitations of the actual capacity of satellite images, especially for yield estimation. We shall present a brief overview of the MARS Project to make up for the few existing references in statistical journals (Ambrosio 1993, Gallego 1992). Other presentations can be found in conference papers (Meyer Roux 1990, Delincé 1990, Sharman *et al.* 1992, Carfagna *et al.* 1994) or remote sensing journals (González *et al.* 1991, Gallego *et al.* 1993).

2. THE MARS PROJECT OF THE EUROPEAN COMMUNITY

The MARS Project was launched in 1988 to assess and to develop operational applications of Remote Sensing to Agricultural Statistics. It is carried out by the Institute of Remote Sensing Applications (IRSA) of the Joint Research Centre (JRC) of the EC. Most of the activities of the period 1988-1993 were divided into 4 main parts, named "actions":

(1) Regional Crop Inventories.
(2) Monitoring Vegetation.
(3) Agrometeorological Models.
(4) Rapid Estimates at the EC level.

Some work is made as well in other related fields, such as area frame sampling. We shall focus here on a sampling method used in the frame of action 1 "Regional Inventories", but we shall first say a word about the other actions.

2.1 Monitoring Vegetation

This action deals with low resolution satellite images from NOAA-AVHRR (Advanced Very High Resolution Radiometer). In these images each pixel has about 1 km^2 in the vertical of the satellite orbit. The main objectives are the development of friendly software for the pretreatment of these images, and building a data bank with time series vegetation indexes and other indicators for about 3,000 monitoring units in the EC. These monitoring units have not yet been definitely defined. They should be geographic areas roughly between 500 km^2 and 1,000 km^2 with a more or less homogeneous vegetation or greenness index (Houston 1984, Goward 1991).

2.2 Agrometeorological Models

General and crop specific models are being currently developed on the basis of data from a network of about 650 Meteorological Observatories in Europe and surrounding areas. This model CGMS (Crop Growth Monitoring System), developed in collaboration with the WOFOST (World Food Studies Centre, in Wageningen, Netherlands), also uses other data, such as soil and elevation data, together with information on the physiology of plants (van Diepen 1989, van Lanen 1992). Remote sensing (low resolution images) will come into the picture later for the spatial interpolation of ground observed meteorological data. Parameters of the model are currently computed for each cell of a 50 km × 50 km grid.

[1] F.J. Gallego, Joint Research Centre of the European Communities, tp. 440, 21020 Ispra, Varese, Italy; J. Delincé, Commission of the E.C. DG VI, Loi 120, 4-23, 1049 Brussels, Belgium; E. Carfagna, Department of Statistics, University of Bologna, V. Belle Arti 41, 40126 Bologna, Italy.

2.3 Rapid Estimates at the E.C. Level

The main goal is giving rapid estimates of area and yield change of annual crops compared with the previous year based on a two-stage sampling scheme: 53 sites (Figure 1) of 40 km × 40 km with a sample of 16 squared segments of 700 m × 700 m (Figure 2) in each of the sites. Individual data are acquired by photo-interpretation of SPOT-XS or Landsat-TM images. An average of three images is analysed for each site with a minimum of ground information, namely a general knowledge of the dominant crops in each area. A ground survey is made for an a posteriori validation of the photo-interpretation. A monthly report (from March to November) is produced with an update of the estimates. Each report should use all the images acquired more than 15 days before.

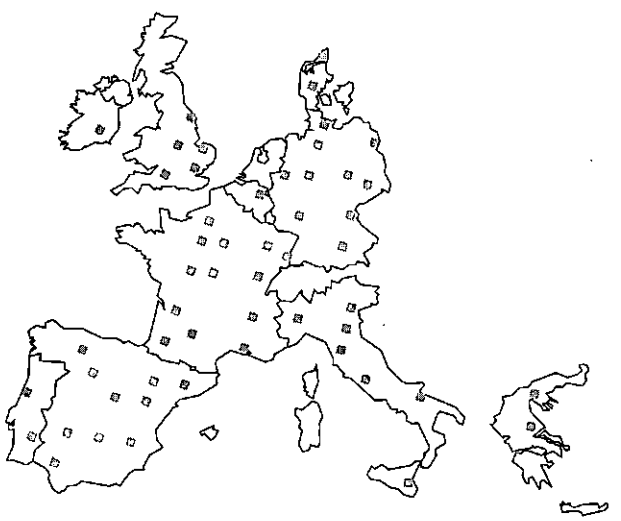

Figure 1. Sample of 53 sites for rapid crop estimates in the E.C.

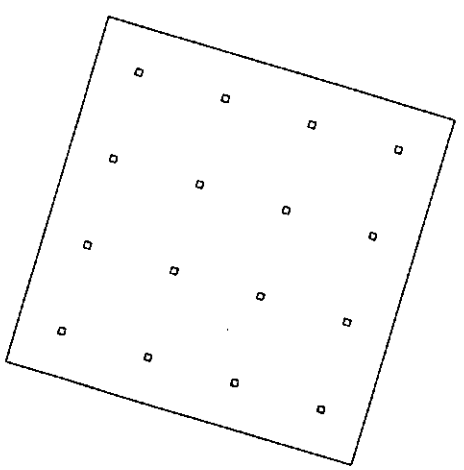

Figure 2. Segments in one site (rapid estimates in the E.C.)

2.4 Regional Crop Inventories by Segment Sampling and Remote Sensing

The objective of the action was to implement, to adapt and to assess estimation methods for crop area and production based on area frame sampling and satellite images. When this action was implemented by the IRSA in 1988 on five pilot regions of approximately 20,000 km^2 each; an absolute priority was given to annual crops: soft and durum wheat, barley, rapeseed, dried pulses, sunflower, maize, cotton, tobacco, sugar beet, potatoes, rice and soya, as well as fallow. Attention is being shared more and more by permanent crops, pastures, and non-agricultural land uses.

Since 1990 the IRSA has progressively transferred the initiative to regional or national administrations that wish to use area frame surveys based on segments. In general, the activities have been shifted to the southern countries of the EC and the former communist countries in central Europe, that have shown much interest in the method (Figure 3). In some cases, like in Italy, there is just an exchange of points of view between the national project and the IRSA.

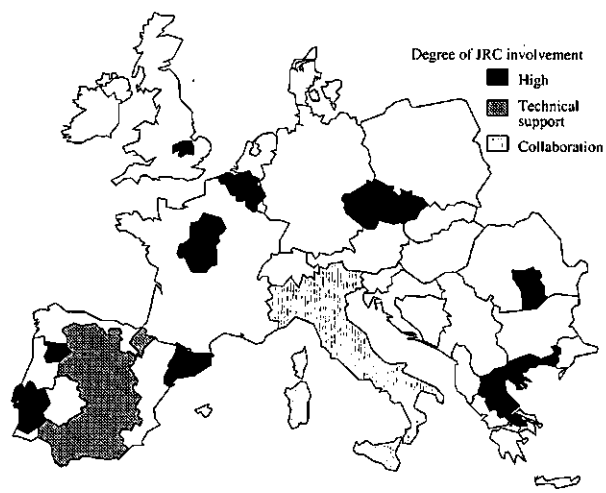

Figure 3. European regions with segment surveys in 1992.

2.4.1 Sampling Segments on a Square Grid

There are two main approaches to building an area frame based on segments: the segments can be drawn on topographic or cadastral maps following roads, rivers, or limits of fields (sometimes called cadastral segments). The sample is usually drawn with a two-stage procedure with intermediate primary sample units to reduce the burden to build the frame (Cotter 1987), which remains in any case a heavy operation.

We generally use frames based on a square grid (Gallego and Delincé 1994), which is much faster to define. Satellite images are generally (but not necessarily) used for stratification prior to sampling.

Figure 4 illustrates a small example of this kind of sample with a very simple stratification and segments of 25 ha (hectares). Sampling is systematic, repeating a pattern in square blocks. In this case the blocks have a size of 10 km × 10 km, and the pattern has 4 replicas in the most agricultural stratum (plain), 2 replicas in the hills, and one in the mountains.

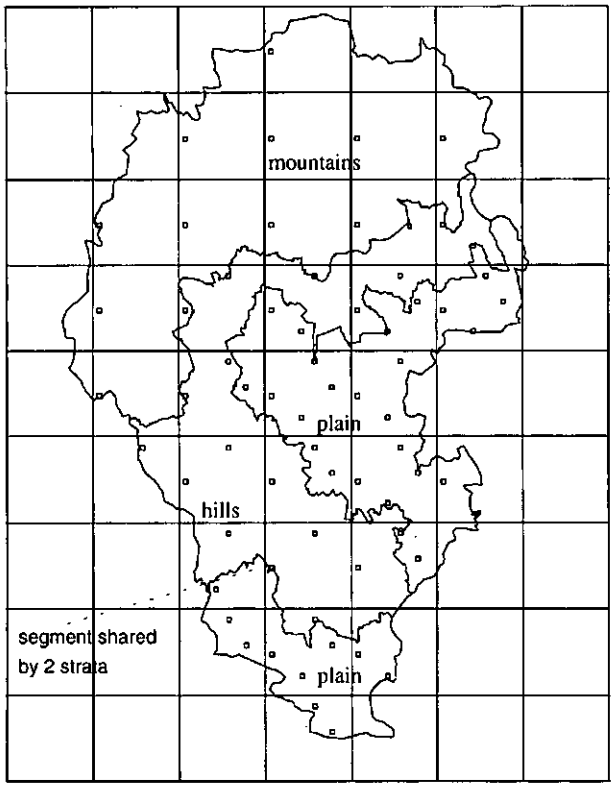

Figure 4. Example of area frame sample with squared segments and squared blocks.

The main drawback of this approach is the management of segments that fall on the boundary between two strata (Figure 5). Three alternatives are being tested for this problem: (1) adapting the stratification to the sampling grid, (2) splitting border segments into pieces that belong to different strata, and (3) keeping only the largest one among these pieces.

The most frequent non-sampling errors – shifts in location and inaccuracy in shape or size of the segment – are not strongly correlated with land use. No major influence has been found on the area estimates or their precision.

The sample pattern to be repeated in each block is drawn at random with a restriction on the distance between segments in order to avoid segments that are too close to each other. Cluster estimators can be used in this case rather than standard formulae for random sampling (Fuentes 1994, Ambrosio 1993). Systematic sampling has a risk of bias if there is a cyclic effect in the landscape with a period that coincides with the block size (10 km in the example), but this is very unlikely. The distance threshold between segments can induce an overestimation of standard errors if the spatial correlation is significantly positive for distances less than the threshold.

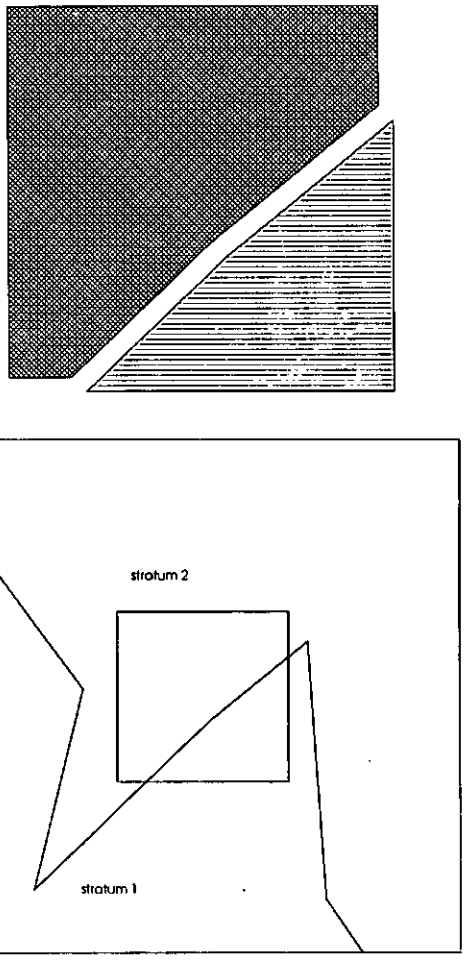

Figure 5. A segment can be split by a stratum boundary.

The size of the segments varies from region to region depending on the agricultural landscape, especially on the size of fields. In the Czech Republic, the segment size was 400 ha. For the area survey, enumerators locate the segments, draw fields on a transparent sheet placed over an aerial photograph, and write down their land use. About 5% to 10% of the segments are visited again by supervisors to check for possible errors on the ground work. Satellite images are not used either for the survey itself or for the farm survey, but they can be optionally used to improve the precision of the area estimates as described in the next section.

2.4.2 Improving Area Estimates with Satellite Images

High resolution satellite images from Landsat-TM or SPOT-XS sensors have been assessed and are still being used at moderate scale to improve the estimates obtained from the ground survey on a sample of segments. The most commonly used approach is the regression estimator on classified images. An alternative estimator based on confusion matrices has been tested with results that are very close to those of the regression estimator (Hay 1988, Gallego 1994).

The conclusions of this assessment are similar to those of the US Department of Agriculture (Allen 1990). The use of satellite images for area estimation is operational, but still too expensive for the efficiency obtained. The economic threshold can be reached by improving image processing automation, since the cost of image processing in the European market for this purpose is much higher than the cost of the images themselves. This threshold has nearly been reached with Landsat-TM images in Greece. Different conclusions on cost analysis are presented by Giovacchini (1992).

3. SAMPLING FARMS BY POINTS

For agricultural surveys in the European Community, farms are traditionally sampled from a list frame (Eurostat 1991). The list is a census of farms that exceed a certain size threshold. In many countries an agricultural census is made every 10 years and is seldom updated (if ever). Hence there may be a substantial difference between the sampling frame and the actual population at the date of the survey. The situation is worse in the central European countries of the former Eastern Block (the area between Poland and Rumania-Bulgaria), where the change of land property structure is so rapid that the census may not exist for private farms and becomes obsolete for co-operatives.

Area frames on square segments can be easily defined when the geographic borders of the region are known. A subsample of these segments is used as well for sampling farms in several countries with the help of a template of points overlaid on the segment. This has been experimentally tested in Germany, Portugal, Italy (Carfagna 1991) and Spain, and is now being regularly used in Greece, Rumania and the Czech Republic.

The template is the same for all the segments in a stratum, and usually symmetric to reduce the risk of bias due to a particular geographic location. Data are obtained only for farms corresponding to points falling on Utilized Agricultural Area (UAA).

The definition of UAA used in the field work is adapted to each national system. Farm buildings and rough pastures are included in some countries and excluded in other countries. The crucial point is that the definition used must be consistent with the definition of the column UAA used for computation (Table 1).

Table 1

Observations Generated by Points Sampled in the Segment of Figure 6

Segment	Point	UAA	Permanent Crops	Wheat		Barley	
				Area	Production	Area	Production
1	1	19	4	12	64	0	0
1	2	0	0	0	0	0	0
1	3	0	0	0	0	0	0
1	4	35	0	24	131	3	12
1	5	35	0	24	131	3	12
2	...	...	...	...	...	...	...

In the example of figure 6, point 3 fell on woodland and point 2 on a built area. They will generate two zero-valued records in the farm file. The enumerator will have to locate the farmers for the other three points. The farm corresponding to point 1 has other fields in the segment, that will be implicitly included in the survey, but the enumerator will not need to find out if these fields exist. Points 4 and 5 belong to the same farm, and it will appear twice in the farm file (Table 1).

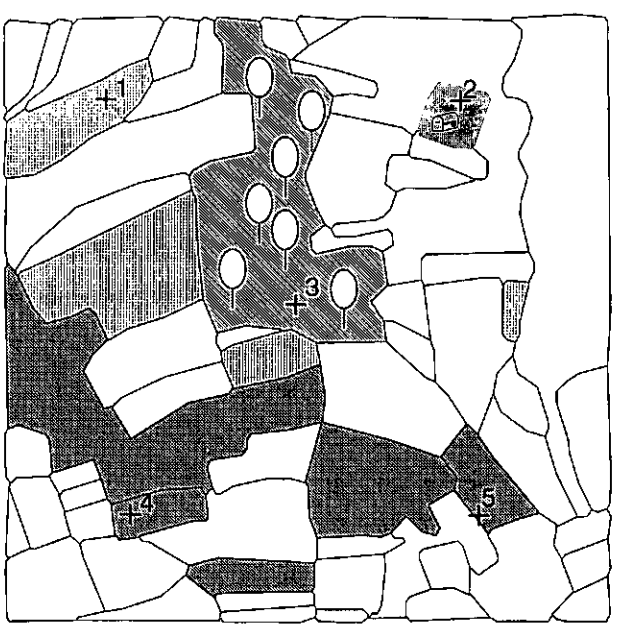

Figure 6. Segment with a pattern of 5 points for farm sampling.

Farmers are located and asked to provide global data for the farm, including total area and production of each target crop. No question is asked about the production of each field or the set of fields inside the segment. This is not necessary because in the final formulae to compute the estimates (formulae 2 and 3 in section 4.1) the crop area or the production in the tract is not used.

The ground survey instructions are usually transferred from the JRC to National Administrations. They explain the instructions to Regional co-ordinators, who give the information to the enumerators. Instructions may be modified in some of these steps. Checking that the instructions have not been misunderstood is sometimes difficult, in part because linguistic limitations are a serious barrier to direct contact with enumerators. In some countries (*e.g.*, Spain) farmers live mainly in rather large urban nuclei and are difficult to locate; this can lead to a significant amount of missing data.

4. ESTIMATES BASED ON FARMS SAMPLED BY POINTS

We assume that the population Ω of segments is divided into strata Ω_h, $h = 1, \ldots, H$, the total population size is N segments (N_h for stratum Ω_h) and the sample size is n segments (n_h). The size of our sample of points in each segment will be K_i, previously fixed; in general we have $K_i = K$, constant across all strata, out of which F_i correspond to the farms on which these points fall. Each segment i has a total UAA surface U_i.

We have a two-staged sampling scheme. In the first stage the segment i is selected with probability $p_i = 1/N_h$ in each of the n_h trials. In the second stage the unit is not the farm but the tract (UAA in a segment, that belongs to the same farm). The tract k of segment i has an area T_{ik}. The total UAA of the farm is A_{ik} over all segments. U_i is the sum of the tracts T_{ik} in segment i.

The method presented below is closely related to the so called "weighted segment estimator" approach used in the U.S. and in Canada (Nealon 1984).

4.1 Estimates Based on Farms and Non-Farm Points

There will be $K - F_i$ observations (fictitious farms) with value 0 corresponding to points outside the UAA.

Sampling through points means that tracts are selected with replacement and with a probability p_{ik} proportional to the area T_{ik}/D_i, (the knowledge of T_{ik} is not necessary), where D_i is the size of the segment determined by the frame design. We are implicitly assuming that the surveyed region is flat. A slight bias might be introduced by the fact that annual crops are usually on more or less flat land and pastures or non-UAA are often on land with a steeper slope.

The sampling is done with replacement: a farm can be selected more than once, which gives easier formulae for variance estimation. Strictly speaking the joint selection probability that farms k and k' are in the sample $p_{ikk'} \neq p_{ikk} \times p_{ikk}$ as would be the case if the different points of the template were drawn independently, since there is usually a relatively large distance between them. We will disregard this fact in this paper.

W_{ik} will be an additive quantity for a farm, most often the production or the area of a particular crop. It is obvious that yield is not an additive variable.

Since we have no information about how W_{ik} is distributed inside the farm, we create a fictitious variable X that is uniformly distributed, and that has, by definition, the same total as W for each farm:

$$X_{ik} = \frac{T_{ik}}{A_{ik}} W_{ik}. \tag{1}$$

Estimating the totals of X and W are equivalent problems.

The two-stage version of the Horvitz-Thompson estimator for the total of X in the stratum Ω_h gives:

$$\hat{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{X_{ik}}{p_{ik}} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{D_i}{K_i} \sum_{k=1}^{K_i} \frac{W_{ik}}{A_{ik}}. \tag{2}$$

This means that, even if the second stage sampling unit is the tract, we do not need to know its area nor X_{ik}, but just the global information about the farm.

The estimator is a linear function of the estimates on the selected segments. Its variance in stratum Ω_h can be estimated as (Cochran 1977, section 11.6):

$$\hat{V}(\hat{X}_h) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \sum_{i=1}^{n_h} \frac{(\hat{X}_i - \bar{X}_h)^2}{n_h - 1} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{K_i(K_i - 1)} \sum_{k=1}^{K_i} \left(\frac{W_{ik} D_i}{A_{ik}} - \hat{X}_i\right)^2. \tag{3}$$

The estimates for the total are:

$$\hat{X} = \sum_{h=1}^{H} \hat{X}_h \qquad \hat{V}(\hat{X}) = \sum_{h=1}^{H} \hat{V}(\hat{X}_h). \tag{4}$$

Crop areas are currently estimated from the segment survey with more objective ground data (direct observation of the enumerator on the filed), although some bias can appear due to the imperfect location of the segments on the ground. Farm surveys provide both area and production estimates, but they can have more significant bias due to non response and to a subjective tendency of the farmer that can depend on whether he is more concerned about taxes or about subsidies at the time of the survey. Comparing both area estimates, from segment survey and farm survey, can be useful to check for possible bias on the production estimate based on the farm survey.

The estimates are also possible for cattle, but the results will be presumably bad if there are a substantial number of farms without any UAA, which will not be sampled: the coverage of the area frame will not be complete in this case. On the other hand it may happen that the number of livestock does not correlate to the UAA and hence to the probability of selection. This results in inefficient estimates.

A program in C for Personal Computers has been written (Dicorato 1993) to compute estimates using this method. The main part of the program was first written to compute estimates on a segment survey.

4.2 Estimation Based Only on Farm Points

We shall mention another option that consists of using only points that fall in the UAA. In this case, we first fix F_i, the number of points that fall in UAA (often $F_i = F_h$, constant in each stratum). In segment i we observe as many points as necessary to have F_i points in the UAA. If the segment i has no UAA, one observation (fictitious farm) is added with 0 values. This is actually an implicit second-stage stratification or stratification of the first-stage units (segments) into two strata; UAA and non-UAA. The non-UAA stratum is not sampled. In this case (2) and (3) are to be adapted substituting K_i by F_i and D_i by U_i. Some inconsistency may arise in hilly areas because A_{ik} comes from the farmer's declaration and U_i from segments drawn on the ground over aerial photographs.

$$\hat{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h} \sum_{h=1}^{n_h} \frac{1}{F_i} \sum_{k=1}^{K_i} U_i \frac{W_{ik}}{A_{ik}}, \qquad (5)$$

$$\hat{V}(\hat{X}_h) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \sum_{i=1}^{n_h} \frac{(\hat{X}_i - \bar{X}_h)^2}{n_h - 1} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{F_i(F_i - 1)} \sum_{k=1}^{K_i} \left(\frac{W_{ik} U_i}{A_{ik}} - \hat{X}_i\right)^2, \qquad (6)$$

the second term of (6) is null for segments with no UAA. This term cannot be computed if $F_i = 1$ because of non-response. A value 0 can be attributed, though this will lead to an underestimation of the within-segment variance, which is relatively small according to calculations made on available data (Carfagna 1992).

This approach has only been used once to resolve a misunderstanding of instructions for ground work that should have been performed following the method in section 4.1. However advantages and drawbacks of both approaches are not clear, and no systematic comparison has been made so far on the same region and same year. Using only farm points can increase the cost of the survey if the number of points per segment is to be kept constant, but the non-UAA points removed correspond to null values of W_{ik}, and their removal can result in a reduction of the variance.

4.3 Farms with Fields in Different Strata

At first sight, the estimator (2) seems to assume that a farm k that has been selected through a point in stratum Ω_h is completely included in this stratum. It is obvious that a farm can have fields in different strata, and the question arises as to whether this fact disturbs the reliability of the results.

We stress again that the variable used is not really W_{ik}, but X_{ik} defined for each individual tract. The total of W does not coincide with the total of X in each stratum, but it does in the whole region as long as

$$\sum_i T_{ik} = A_k. \qquad (7)$$

Notice that A_k is identical to what we have called previously A_{ik}, where the subindex i is used only to indicate that farm k has been selected in the sample through segment i.

This identity holds on the population, regardless of the sampling procedure, if the farms are entirely inside the region and if the geometry of the ground survey document (aerial photograph) is correct.

The perturbation due to farms with fields in different regions is expected to be small because of the low proportion (generally under 1–2%) and because there is a compensation between the bias due to fields inside the region belonging to farms with the headquarters outside the region and vice versa. We are assuming that the total of W is calculated on the farms that have their headquarters inside the surveyed region.

4.4 Nonresponse

We refer here to the estimators based on farm and non-farm points (section 4.1). If a farmer does not co-operate or cannot be found, the corresponding row or rows of the input table (Table 1) are substituted with the average values of responding farms in the segment, if there are any; otherwise they are substituted with the average of responding farms for all the segments in the current stratum.

If in the second stage (sampling farms inside the segment) we consider farm and non-farm points, and give value 0 to the points that fall in non agricultural land, it is obvious that the exclusion of nonrespondents would produce a serious bias, because the zero values corresponding to non-UAA are never missing. These points are not used to compute the "average farm" values used to fill missing values. There is still a risk of bias if farmers who cannot be located or refuse to co-operate have a peculiar behaviour, *e.g.*, if they are on the average smaller or less efficient farms.

We could have considered a different way of overcoming this problem: eliminating both missing values and a proportional number of 0 values corresponding to non-UAA points. Both give the same estimate for the total, but the second solution is more uncomfortable because the sample size in the second stage is not an integer any more.

The introduction of "average farm" values will lead to a negative bias on the variance. To compensate it, the farm is not included in the sample size K_i for the computation of variances.

5. RESULTS: TWO EXAMPLES

We discuss below some results from two regions: Emilia Romagna (Italy) and the Czech Republic. In the Czech Republic, the method presented in sections 2.4, 3 and 4.1 was used; there were no missing data at all. In Emilia Romagna the general design of the survey did not follow exactly the procedure outlined above. Missing data were treated as stated in section 4.4.

5.1 Emilia Romagna 1990

In Emilia Romagna an area of 19,500 km^2 was divided into 4 strata, excluding mountainous areas. A sample of 313 "cadastral" segments (with physical boundaries) was drawn based on a two-staged procedure with primary sampling units (psu) of about 10 km^2. Segment size was approximately 50 hectares or 100 hectares, depending on the strata. 5 points per segment were drawn at random from a grid with a 50 metre step.

Out of the 1,565 points sampled: 326 were non-UAA, the farmer's address could not be found for 206 UAA points, 38 farmers were not located and 32 refused to co-operate. 963 UAA points from 285 segments had valid data, corresponding to 617 farms, some of which appear more than once in the sample.

When we think only of area estimation, the segment survey can be seen as more objective and complete, since there are no missing data and observations do not rely on farmers' answers. If we accept this principle we can have an idea of a possible bias in the farm survey by comparing with the area estimates of the segment survey. Estimates can be compared in Table 2 for the main crops in the region. Figures match well for cereals, excepting durum wheat, and permanent crops, but some problems appear for sugar beet and soya, that might be related to misunderstandings on how to declare second crops in the same year and the same field, or with a bias due to missing values. Official statistics are produced taking into account a variety of information. Durum wheat is reported separately because of the special meaning of this crop due to the significant subsidy granted by the EC to each hectare of crop.

Table 2

Results of the Segment Survey and the Farm Survey for Main Crops in Emilia Romagna (1990)

Emilia Romagna	Segment Survey		Farm Survey				ISTAT
	Area × 1,000 ha		Area × 1,000 ha		Prod. × 1,000 tm		Area
	Esti-mation	CV %	Esti-mation	CV %	Esti-mation	CV %	
Soft wheat	212	5.7	208	6.9	1,177	8	212
Durum wheat	46	14.9	48	15.2	260	14	72
Barley	43	11.2	50	17.7	184	17	38
Rice	-	-	4	59.0	23	61	6
Sugar beet	111	7.1*	96	9.6	5,474	28	119
Soybeans	76	6.0*	55	11.6	321	39	47
Vineyards	78	13.3*	76	18.7			75
Orchards	91	13.1*	96	19.7			85

* Estimate corrected by regression on classified satellite image.
ISTAT: Official statistics. No precision provided.

The coefficients of variation in the farm survey have a reasonable behaviour for cereals, but become more difficult to understand for sugar beet and soybeans. The high CV (Coefficient of variation) for the production can be due to higher yields in larger, more specialized farms.

A correction of the production estimate can be made using the difference of area estimates between the segment survey and the farm survey. A regression estimator approach might be a good solution.

Livestock is seriously underestimated (Table 3) since many livestock owners do not have agricultural land. A mixed approach was used for cattle and pigs with an exhaustive survey using a list frame for the 50 largest farms and point sampling for the rest. The procedure works for pigs, but CVs are not yet satisfactory.

Table 3

Results of the Farm Survey on Area Frame and Mixed Frame for Livestock in Emilia Romagna (1990)

× 1,000 Units	Census	Area Frame		Mixed Frame	
		Estimate	CV %	Estimate	CV %
Cattle	869	829	14	894	13
Pigs	1,876	1,312	37	1,818	27
Sheep	90	38	74		

5.2 Czech Republic 1992

Area frames seem especially useful in the former communist countries in Europe because of the rapid change of property structure. Agricultural statistics are mainly produced with no sampling error by adding the data reported by each state farm or co-operative. This procedure will collapse in the coming years. It will be extremely

difficult to have an idea of the number of existing farms, and an agricultural census will be out of date before the data are elaborated. Area frames might be the best alternative.

The territory of the Czech Republic (about 80,000 km^2) has been stratified into 6 strata by photo-interpretation of Landsat-TM images. The stratification needed 15 working days for one person. In 1992, a survey was made with a sample of 417 square segments of 400 ha drawn by repetition of a fixed pattern on blocks of 40 km × 40 km. Segments were visited and area estimates obtained as explained in section 2.4.1.

Farms have been sampled using a fixed grid of 5 points in each segment. The shape of the 5-point grid was in "x" like in figure 6. This procedure gave 2,085 points: 858 non-agricultural, and the other 1,227 from 458 farms. No missing data were recorded: all the farms were identified and none refused to co-operate. This happened mainly because the old structure of large farms was still nearly intact.

Table 4 compares the results of the segment survey (direct observations on the field), the farm survey (farms sampled by points), and official statistics for the main crops in the country. Official statistics are obtained by adding figures reported by all the state farms or co-operatives. There is a moderate disagreement on area estimates for wheat, maize, and potatoes. We should not exclude a bias in farmers' answers that has to do with self-consumption of agricultural products.

Table 4

Results of the Segment Survey and the Farm Survey in the Czech Republic (1992)

× 1,000 ha	Segment Survey		Farm Survey				CSO	
	Area	CV%	Area	CV%	Prod.	CV%	Area	Prod.
Wheat	824	5.4	757	3.7	3,412	4.9	780	3,413
Barley	655	5.1	630	3.8	2,521	4.3	640	2,512
Rapeseed	140	11.6	137	6.8	310	7.5	136	296
Sugar beet	119	11.5	127	8.1	4,172	11.0	125	3,874
Maize	361	7.5	326	4.8	8,884	4.3	361	8,904
Potatoes	109	13.6	92	7.9	1,706	8.7	111	1,969

CSO: Czech Statistical Office.

The coefficients of variation (CV) of the area estimates are lower in the farm survey than in the segment survey. This is not surprising since the farm survey gives information about fields outside the segments. The 458 selected farms represent more than 15% of the total UAA in the country. The CVs for production estimates are slightly higher than for area estimates (even lower in the case of maize). This seems to indicate that the variability of yields contributes less than the variability of areas to the variability of production.

6. CONCLUSIONS AND RECOMMENDATIONS

Area frames based on square grids are a pragmatic alternative to area frames based on ground elements delimited by physical features. They are much cheaper to build and they do not seem to have major drawbacks regarding the final results. However some theoretical work is still needed to determine under which conditions the location errors due to non-physical limits have a negligible effect on the estimates.

Sampling points inside area segments provides a feasible way to build frames for farm sampling. They are extremely useful if list frames (census) are poorly updated or do not exist. Sampling a few points per segment can be much cheaper than surveying all the farms with fields in the segment. Five points per segment seems to be a reasonable choice.

Area frames alone give poor results for livestock when the number of units is not strongly correlated with Utilized Agricultural Area of the farm.

ACKNOWLEDGEMENTS

We are grateful to the various National and Regional Administrations that have collaborated for this work, and to A. Burrill and O. O'Hanlon for the kind revision of the paper. The numerous comments by the referees have been essential for the paper to become more useful to readers.

REFERENCES

ALLEN, J.D. (1990). A look at the remote sensing applications program of the national agricultural statistics service. *Journal of Official Statistics*, 6, 393-409.

AMBROSIO, L., ALONSO, R., and VILLA, A. (1993). Estimación de superficies cultivadas por muestreo de áreas y teledetección. Precisión relativa. *Estadística Española*, 35, 91-103.

CARFAGNA, E., RAGNI, P., ROSSI, L., and TERPESSI, C. (1991). Area frame: un Nuovo Istrumento per la Realizzasione delle Statistiche Agricole in Italia. *Contributi alla Statistica Spaziale*. University of Parma.

CARFAGNA, E., and DELINCÉ, J. (1992). Farm survey based on area frame sampling. The case of Emilia Romagna in 1990. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publication of the E.C. Luxembourg.

CARFAGNA, E., and GALLEGO, F.J. (1994). Extrapolating intra-cluster correlation to optimize the size of segments in an area frame. Conference on Applied Statistics to Agriculture, Kansas State University, Manhattan, KS.

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

COTTER, J., and NEALON, J. (1987). Area frame design for agricultural surveys. U.S. Department of Agriculture. National Agricultural Statistics Service.

DELINCÉ, J. (1990). Un premier bilan de l'action 1 Inventaires Régionaux du Projet Agriculture après deux années d'activité. Conference on the Application of Remote Sensing to Agricultural Statistics, Varese. Office for Publication of the E.C. Luxembourg.

DICORATO, F. (1993). AIS estimation programs. User documentation. JRC Ispra.

VAN DIEPEN, C.A., WOLF, J., VAN KEULEN, H., and RAPPOLDT, C. (1989). WOFOST: A simulation model for crop production. *Soil Use and Management*, 5, 16-24.

EUROSTAT 1991. Working party, Crop Products Statistics. Methodological reports. Document AGRI/PE/333, Luxembourg.

FUENTES, M., and GALLEGO, F.J. (1994). Stratification and cluster estimator on an area frame by squared segments with an aligned sample. Conference on Applied Statistics to Agriculture, Kansas State University, Manhattan, KS.

GALLEGO, F.J. (1992). Flächenschatzungen für einjährige Feldfrüchte mit Hilfe Fernerkundung. *Neue Wege raumbezogener Statistik. Forum der Bundesstatistik*, 20, 109-120. Wiesbaden: Statistisches Bundesamt.

GALLEGO, F.J. (1994). Using a confusion matrix for area estimation with remote sensing. *Atti Convegno AIT*, Roma, 99-102.

GALLEGO, F.J., and DELINCÉ, J. (1994). Area estimation by segment sampling. In *Euro-Courses Remote sensing applied to Agricultural Statistics*.

GALLEGO, F.J., DELINCÉ, J., and RUEDA, C. (1993). Crop area estimates through remote sensing: Stability of the regression correction. *International Journal of Remote Sensing*. 14, 18, 3433-3445.

GIOVACCHINI, A. (1992). Agricultural statistics by remote sensing in Italy: an ultimate cost analysis. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publication of the E.C. Luxembourg.

GONZÁLEZ, F., LOPEZ, S., and CUEVAS, J.M. (1991). Comparing two methodologies for crop area estimation in Spain using landsat TM images and ground gathered data. *Remote Sensing of the Environment*, 32, 29-36.

GOWARD, S.N., MARKHAM, B., DYE, D.G., DULANEY, W., and YANG, J. (1991). Normalized difference vegetation index measurements from the advanced very high resolution radiometer. *Remote Sensing of the Environment*, 35, 257-277.

HAY, A.M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9, 1395-1398.

HOUSTON, A.G., and HALL, F.G. (1984). Use of satellite data in agricultural surveys. *Communications in Statistics Theory and Methods*, 13, 23, 2857-2880.

VAN LANEN, H.A.J., VAN DIEPEN, C.A., REINDS, G.J., DE KONING, G.H.J., BULENS, J.D., and BREGT, A.K. (1992). Physical land evaluation methods and GIS to explore the crop growth potential and its effects within the European Communities. *Agricultural Systems*, 39, 307-328.

MEYER-ROUX, J. (1990). Présentation du projet pilote de télédétection appliquée aux statistiques agricoles. Conference on the Application of Remote Sensing to Agricultural Statistics. Office for Publications of the E.C. Luxembourg.

NEALON, J.P. (1984). Review of the multiple and area frame estimators. U.S. Department of Agriculture, Statistical Reporting Service, Report 80, Washington, D.C.

SHARMAN, M., and de BOISSEZON, H. (1992). Action IV: de l'image aux statistique, bilan opérationnel après deux années d'estimations rapides des superficies et des rendements potentiels au niveau Européen. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publications of the E.C. Luxembourg.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 117–124
Statistics Canada

Use of Capture-Recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable

K.H. POLLOCK, S.C. TURNER and C.A. BROWN[1]

ABSTRACT

We present a formal model based sampling solution to the problem of estimating list frame size based on capture-recapture sampling which has been widely used for animal populations and for adjusting the US census. For two incomplete lists it is easy to estimate total frame size using the Lincoln-Petersen estimator. This estimator is model based with a key assumption being independence of the two lists. Once an estimator of the population (frame) size has been obtained it is possible to obtain an estimator of a population total for some characteristic if a sample of units has that characteristic measured. A discussion of the properties of this estimator will be presented. An example where the establishments are fishing boats taking part in an ocean fishery off the Atlantic Coast of the United States is presented. Estimation of frame size and then population totals using a capture-recapture model is likely to have broad application in establishment surveys due to practicality and cost savings but possible biases due to assumption violations need to be considered.

KEY WORDS: Incomplete frames; Capture-recapture sampling; Angler surveys; Telephone surveys; Access surveys.

1. INTRODUCTION

In classical sampling theory it is assumed that a complete frame exists. There is, at least conceptually, a complete list of population units. It is then possible to draw a probability sample from the population. Estimators of population parameters such as mean or total then have known properties and are easily studied theoretically or numerically. Books on sampling theory such as Cochran (1978) concentrate on this situation and give properties of estimators for common sampling designs such as simple random sampling, stratified random sampling and multistage (cluster) sampling.

In practice in surveys of establishments or businesses a complete frame may not exist. Lists of establishments kept by professional associations or government agencies are often incomplete. One approach to tackling this problem is to use the multi-frame approach originally developed by Hartley (1962, 1974). Examples of this approach are the National Agricultural Statistics Service (USDA) farm surveys (Vogel and Kott 1993). These surveys use an incomplete list frame of farms plus an area frame where all farms within a sample unit are enumerated. Therefore the list frame is incomplete while the area frame is conceptually complete. (There is a list of all area units and within each area unit theoretically all farms could be enumerated.)

There are some situations, however, where it may not be possible to use an area frame for practical reasons. All that the researcher may have available may be several incomplete list frames of establishments. The usual approach in this situation is to merge all the incomplete lists and ignore any remaining incompleteness. Depending on the degree of incompleteness remaining there could be serious negative bias on estimates of population size and population total.

Later we present a formal model based sampling solution to this problem based on capture-recapture sampling. Capture-recapture sampling models are widely used in sampling animal populations (Seber 1982) and also for adjusting the U.S. census for undercoverage (Feinberg 1992). In the simplest case of two incomplete lists we consider "marked" units to be those which occur on both lists and unmarked units to be those which do not occur on both lists. It is easy to estimate total frame size using the Lincoln-Petersen estimator (Seber 1982, p. 59). This estimator is model based with a key assumption being independence of the two lists. Once an estimator of the population size has been obtained it is possible to obtain an estimator of population total for some characteristic if a sample of units has that characteristic measured.

The usual estimator of a population total for simple random sampling without replacement is

$$\hat{Y} = N\bar{y}, \tag{1.1}$$

where N is known and $\bar{y}$ is the mean of the sample, see for example Cochran (1978, p. 21). The variance of $\hat{Y}$ is given by

$$\mathrm{Var}(\hat{Y}) = N^2\mathrm{Var}(\bar{y}), \tag{1.2}$$

[1] K.H. Pollock, North Carolina State University, Raleigh, NC 27695; S.C. Turner and C.A. Brown, National Marine Fisheries Service, Miami, FL 33149, U.S.A.

where

$$\text{Var}(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right),$$

S^2 is the population variance and $(N - n/N)$ is called the finite population correction factor. The estimator (1.1) is also an unbiased estimator of the population total.

Here our estimator is

$$\hat{Y} = \hat{N}\bar{y}, \tag{1.3}$$

where $\hat{N}$ is obtained from the capture-recapture method.

This means the properties of the estimator (1.3) are more difficult to evaluate because both $\hat{N}$ and $\bar{y}$ are random variables unlike in estimator (1.1) where N is a known quantity. The estimated variance of $\hat{Y}$ here is given by

$$\widehat{\text{Var}}(\hat{Y}) = (\hat{N})^2\,\widehat{\text{Var}}(\bar{y}) + (\bar{y})^2\,\widehat{\text{Var}}(\hat{N}) + \widehat{\text{Var}}(\bar{y})\,\widehat{\text{Var}}(\hat{N}), \tag{1.4}$$

assuming that $\bar{y}$ and $\hat{N}$ are independent and using an exact result due to Goodman (1960). The estimator (1.3) is only an unbiased estimator if $\hat{N}$ and $\bar{y}$ are unbiased estimators of the population size and population mean respectively which is not usually the case in practice. We discuss the estimator (1.3) in the large pelagic fishery survey example in Section 3.

The remainder of the paper is structured as follows. In Section 2 we review the capture-recapture literature to give an overview of the types of models available. In Section 3 we present an example of a sample survey of fishing boats. (We consider a boat analogous to a business establishment). While this example has some unique features we believe it has many features common to other establishment surveys. In the final discussion section we summarize the strengths and weaknesses of using the capture-recapture approach to estimating frame size in establishment surveys. Many of our ideas will require further research.

2. A BRIEF REVIEW OF CAPTURE-RECAPTURE MODELS

It is obviously beyond the scope of this manuscript to review the extensive capture-recapture literature. For more information we recommend Seber (1982), White *et al.* (1982), Pollock *et al.* (1990) and Pollock (1991). Pollock (1991) is a review paper and a good lead into the literature and our treatment in this section follows it very closely. The other references are books and monographs for the serious reader with more time.

Here we briefly discuss the Lincoln-Petersen model for two samples, more general closed population and open population models for more than two samples, and finally a method which combines closed and open population models in one sampling design. Pollock *et al.* (1990, p. 9) presents a flow chart which shows an overview of the models and how they relate to each other.

2.1 The Lincoln-Petersen Model

This is the oldest, simplest and best known capture-recapture model dating back to Laplace, who used it to estimate the population size of France. It was first used in fisheries by Petersen around the turn of the century. An excellent detailed discussion of this model is given by Seber (1982, Chapter 3).

In the original fisheries setting the method can be described as follows. A sample of M fish is caught, marked, and released. Later a second sample of n fish is captured, of which m are marked. An intuitive derivation of the estimator follows from equating the proportions marked in the sample and the population,

$$m/n = M/N, \tag{2.1}$$

which gives

$$\hat{N} = Mn/m. \tag{2.2}$$

A modified estimator with less bias in small samples is due to Chapman (1951) and is given by

$$\hat{N}_c = \left[(M+1)(n+1)/(m+1)\right] - 1. \tag{2.3}$$

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

$$\widehat{\text{Var}}(\hat{N}_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}. \tag{2.4}$$

See for example Seber (1982, p. 60).

The crucial assumptions of this model are:

(a) The population is completely closed to additions and deletions,

(b) all the fish are equally likely to be captured in each sample, and

(c) marks are not lost or overlooked.

The assumption about closure can be weakened, but even for a completely open population, where this estimator does not apply, a modification of the Lincoln-Petersen estimator is used. The assumption of equal catchability causes problems in most applications. There may just be inherent variability (heterogeneity) in capture probabilities of individual animals due to age, sex or other factors. There may also be a response to initial capture (trap response). In the next section, we consider closed

population models with more than two samples that allow for time variation as well as heterogeneity and trap responses in the animals' capture probabilities. The loss or overlooking of marks can be serious. One way to estimate mark loss is to use two marks (Seber 1982, p. 94).

2.2 Closed Population Models

Closed population models require the assumption that no births, deaths, or migration in or out of the population occur between sampling periods. Therefore, these models are generally used for studies covering relatively short periods of time (*e.g.*, trapping every day for 5 consecutive days). Capture histories for every animal caught are the data needed for obtaining estimates under these models. Important early references are Schnabel (1938) and Darroch (1958), who considered models that assumed equal catchability of animals in each sample.

A set of models that allow capture probabilities to vary due to heterogeneity, (h), trap response (b), time variation (t), (*i.e.*, capture probability for time i differs from that for time j) and all possible two- and three-way combinations of these factors is now available. The eight models [$M(o)$, $M(h)$, $M(b)$, $M(bh)$, $M(t)$, $M(th)$, $M(tb)$, $M(thb)$] were first considered as a set by Pollock (1974) and were more fully developed by Otis *et al.* (1978), White *et al.* (1982), and Pollock and Otto (1983). Otis *et al.* (1978) provided a detailed computer program, CAPTURE, for use with their monograph. An updated version provides estimates for seven of the eight models and a model selection procedure that aids the biologist in choosing a model. The model selection procedure is based on a variety of goodness-of-fits tests. Recently, Menkins and Anderson (1988) have emphasized that the model selection procedure is poor for small populations, unless the capture probabilities are unrealistically high.

2.3 Open Population Models

In many capture-recapture studies, it is not possible to assume the population is closed to additions and permanent deletions. The basic open population model suitable for this situation is the Jolly-Seber model (Jolly 1965; Seber 1965; Seber 1982, p. 196). The Jolly-Seber model allows estimation of population size at each sampling time as well as estimation of survival rates and birth numbers between sampling times. Migration cannot be separated from the birth and death processes without additional information.

The Jolly-Seber model requires the following assumptions:

(a) Every animal present in the population at a particular sampling time has the same probability of capture,

(b) every marked animal present in the population immediately after a particular sampling time has the same probability of survival until the next sampling time,

(c) marks are not lost or overlooked,

(d) all emigration is permanent, and

(e) all samples are instantaneous, and each release is made immediately after the sample.

Assumptions (a), (c), and (e) were required under the basic Lincoln-Petersen model described in Section 2.1. Only marked animals are used to estimate survival rates so that, strictly, we do not need to assume equality of marked and unmarked survival rates. In practice however, the biologist will want to use the survival rate estimates to refer to the whole population. The Jolly-Seber model allows for some animals to be lost on capture and hence not returned to the population. The Jolly-Seber model also requires that all emigration is permanent. If animals emigrate and then return to the population this causes so called temporary emigration which is a serious assumption violation and causes major bias in population size estimates.

2.4 Combination of Closed and Open Models

Pollock (1982), Pollock *et al.* (1990) and Kendall (1992) discuss sampling methods which allow the use of closed and open models in one design. One advantage of these methods is that it is possible to allow for unequal catchability whereas in the traditional Jolly-Seber model it is not possible to allow for unequal catchability. They also have the advantage of allowing for temporary emigration of animals.

2.5 Applications of Capture-Recapture Models

Capture-recapture models have obviously been widely applied to wildlife and fishery populations. A variety of novel nonbiological applications of capture-recapture methods have also now appeared. Many authors have applied capture-recapture to estimating the census undercount. (See Feinberg (1992) for a complete bibliography). Cowan, Breakey, and Fischer (1986) used it to estimate the number of homeless people in a city. Greene (1983) has used the method to estimate demographic parameters on criminal populations. Wittes (1974) and Wittes, Colton, and Sidel (1974) have used capture-recapture to estimate numbers of people with illnesses from hospital and other lists. The sampling of elusive human populations using cluster sampling, network sampling, and capture-recapture sampling was discussed by Sudman, Sirken and Cowan (1988).

3. USE OF CAPTURE-RECAPTURE MODELS IN THE LARGE PELAGIC SURVEY

The Large Pelagic survey is an angler survey conducted by the National Marine Fisheries Service using a telephone-access survey design. A sample of fishing boat owners on a list are telephoned to obtain fishing effort (*i.e.*, number

of fishing trips in a period) information. Catch per unit effort (*i.e.*, catch per trip) information is obtained from a second sample of boat owners at access points at completion of their fishing trips. The information from the two surveys is combined to estimate total effort and total catch of important species such as Bluefin Tuna.

A serious problem with this survey is that the list of boat owners used in the telephone survey is very incomplete. Therefore, classical sampling theory which assumes a complete frame of known size (N) is inadequate and has to be modified. The current method of estimating the size of the fishing boat list frame involves combining two lists, (a telephone list with a dockside list) and using the Lincoln-Petersen model. There are questions about whether this is the best approach. For example, it might be possible to combine more than two lists and if so then we could use the closed or open population models reviewed in Sections 2.2 and 2.3. However, we defer those questions and begin by reviewing and evaluating the current method as an example to illustrate the potential usefulness of the approach to other establishment surveys.

3.1 The Lincoln-Petersen Model

3.1.1 Estimation of Frame Size *(N)*

Under the current method the "marked" boats (M) are those on the master list which is primarily derived from previous telephone interviews. The recapture sample is carried out dockside at gas pumps and the total number of boats intercepted (n) is checked to see which ones are "marked" (m) (*i.e.*, on the original master list). Equation 2.3 can then be used to provide an estimator of the frame size (N). Let us now consider the assumptions of this model and what effect violations might have on the bias of the estimator of N.

Closure

This assumption is likely to be violated. Fishing boats may be on the master list and then no longer take part in the fishery (losses). New fishing boats may join the fishery while it is in progress (gains). Ideally a separate estimate of frame size should be obtained for each two week time period. The advantage of using the Lincoln-Petersen closed model estimator is its simplicity and practicality. Biases in the estimator due to lack of closure could be either positive or negative.

Currently it is not known how the fishing fleet size is likely to change during the fishing season. A multiple capture-recapture sampling design would allow use of the Jolly-Seber model to estimate the fleet size during each period. Examination of these estimators and the survival rate and recruitment number estimators will enable us to evaluate the validity of the closure assumption. At the moment we can only make conjectures.

Equal Catchability

Violation of the assumption of equal catchability may be due to either inherent heterogeneity of capture probabilities between individuals or "trap response" where individuals that are marked have higher or lower capture probabilities than unmarked individuals. In either situation when the individuals on the lists are fishing boats we believe there is a potential for heterogeneity of capture probabilities among fishing boats. If heterogeneity is operating across both samples, individuals "caught" on the first list will tend to be those with high capture probabilities and therefore they will more likely to be "caught" again on the second list. This means that the proportion marked in the second sample (list) will be too high and the estimator of N will be negatively biased. Note that this intuitive argument makes clear it is not heterogeneity per se which is the problem but the positive correlation of capture probabilities between the two samples. Another way of stating the equal catchability assumption is that capture probabilities in the two samples are independent. One method of attempting to achieve independence of the capture probabilities in the two samples is to use totally different sampling schemes for the two samples. This is why we recommended earlier that one sample list be based on the telephone interviews and the other on dockside interviews. However, we do suspect that there is still another heterogeneity and lack of independence in capture probabilities. We believe that fishing boats which take a very active part in the fishery are more likely to be on any lists gathered (telephone or dockside). This heterogeneity will cause a negative bias on the estimate of frame size but we have no idea of the degree of this negative bias. A more complete discussion of heterogeneity and independence of samples is given by Seber (1982, p. 86).

Marks Lost or Overlooked

The situation here is a little confusing. At first one might think that in this application there is not a way that a mark could be lost or overlooked. However, this assumes that all boats have distinct names or that if boats do have the same name there is additional information like captain's name which makes all individuals on the lists unique. If there is any problem with lack of uniqueness it may not be clear whether a marked boat has been recaptured or not. Another related point is that agents may make errors in the records which make it hard to match up a recapture with the original record. A standard operating procedure is being developed and documented to minimize these kinds of errors in the future.

3.1.2 Estimation of Total Effort and Total Catch

Total Effort (E) (*i.e.*, the total number of fishing trips taken in a defined period) is estimated by

$$\hat{E} = \hat{N}\bar{e}, \tag{3.1}$$

where $\hat{N}$ is the frame size (Fleet Size) estimate and $\bar{e}$ is the mean fishing effort (*i.e.*, average number of fishing trips taken) obtained from the telephone sample. The evaluation of the properties of this estimator is more difficult than when N is known because both $\hat{N}$ and $\bar{e}$ are random variables. We suspect that $\bar{e}$ is biased high because fishing boats that do not fish much are less likely to be on the list. Unfortunately we cannot say that $\hat{N}$ will always be biased high or low. All three of the assumption violations discussed in 3.1.1 could be important (closure, heterogeneity, and mark loss) and it is not clear what direction the overall bias on $\hat{N}$ would take. The only possible approach is to use simulation with a variety of different scenarios for assumption violations. Using equation (1.4) the estimated variance of $\hat{E}$ is given by

$$\widehat{\mathrm{Var}}(\hat{E}) = (\hat{N})^2 \widehat{\mathrm{Var}}(\bar{e}) + (\bar{e})^2 \widehat{\mathrm{Var}}(\hat{N}) + \widehat{\mathrm{Var}}(\bar{e}) \widehat{\mathrm{Var}}(\hat{N}). \quad (3.2)$$

Total catch (C) is estimated by $\hat{C} = \hat{E}\bar{c}$ where $\hat{E}$ is the estimated total fishing effort and $\bar{c}$ is the average catch per unit effort calculated from the dockside interviews. Properties of this equation are likely to be subject to similar concerns as equation (3.1) and again simulation could be very useful.

3.1.3 Illustration of the Method

In this section we present the frame size estimates and total effort estimates for the Virginia Bluefin tuna fishery in part of 1992. These estimates are a part of a larger survey which covered the east coast of the U.S. from North Carolina to Massachusetts. The estimates are separate for charter boats and private boats.

Frame Size Estimates

Lists of unique private boats and charter boats were compiled mainly by telephone interviews from previous seasons. During the current 1992 season "marked" and "unmarked" boats were captured at gas pumps before or after fishing trips.

For private boats the list size was $M = 335$ boats before the season. A sample of $n = 374$ boats were contacted at gas pumps and of those $m = 49$ were marked. The Chapman estimator is $\hat{N}_c = 2{,}519$, $\widehat{SE}(\hat{N}_c) = 303.08$ and relative $\widehat{SE} = 0.12$.

For charter boats the list size was $M = 47$ before the season. A sample of $n = 31$ boats were contacted at gas pumps and of those $m = 13$ were marked. The Chapman estimator is $\hat{N}_c = 109$ with $\widehat{SE}(\hat{N}_c) = 17.88$ and relative $\widehat{SE} = 0.16$.

Total Effort Estimates

Total effort and total catch were estimated in weekly waves. Here we just illustrate the calculations for the week of the 8th to the 14th of June 1992 for total effort.

Total Effort - Private Boats

$\hat{N}_c = 2{,}519$ boats, $\widehat{\mathrm{Var}}(\hat{N}_c) = 91{,}856.4706$, $\bar{e} = 0.15108$ trips per interview, $\widehat{\mathrm{Var}}(\bar{e}) = 0.001242$ and $\widehat{SE}(\bar{e}) = 0.0352$. Using these estimates we obtain

$$\hat{E} = \hat{N}_c \times \bar{e} = 2{,}519 \times 0.15108 = 380.57 \text{ trips},$$

$$\widehat{\mathrm{Var}}(\hat{E}) = \widehat{\mathrm{Var}}(\bar{e})(\hat{N}_c^2) + \widehat{\mathrm{Var}}(\hat{N}_c)(\bar{e})^2 + \widehat{\mathrm{Var}}(\hat{N}_c)\widehat{\mathrm{Var}}(\bar{e}) = 10{,}091.6633, \text{ and}$$

$$\widehat{SE}(\hat{E}) = 100.45.$$

It is useful to also calculate the variance of total effort assuming that the frame size were known. In this case it is $\widehat{\mathrm{Var}}(\hat{E}) = 7{,}780.9384$ with $\widehat{SE}(\hat{E}) = 88.77$ and this shows that 89% of the standard error of the Total Effort estimate is due to variation in average effort and only 11% is due to estimation of frame size.

Total Effort - Charter Boats

For charter boats $\hat{E} = 59.95$ trips with $\widehat{\mathrm{Var}}(\hat{E}) = 512.5100$ and $\widehat{SE}(\hat{E}) = 22.64$.

The variance of the Total Effort estimate assuming the frame size is known is $\widehat{\mathrm{Var}}(\hat{E}) = 404.8926$ with $\widehat{SE}(\hat{E}) = 20.12$. Again 89% of the standard error of the Total Effort estimate is due to variation in average effort and only 11% is due to estimation of frame size.

3.2 More Than Two Lists

In Section 2 we indicated that there are a lot more modeling possibilities if one has multiple (greater than 2) lists. Here we consider closed and open population models for the more general case. We foresee the sampling scheme as follows. Before the start of the fishing season there would be a preliminary sample to establish a list (either telephone or dockside). During each time period (say two weeks) there would be an additional list compiled using a telephone or dockside survey. Now each individual boat would have a capture history which would indicate which lists it appeared on. (Suppose we have five time periods then a capture history of 1 1 1 0 1 would indicate a boat appeared on the lists in all except the fourth time period).

The structure of the sample and the population would therefore be as in Table 1.The first question that has to be addressed is whether we need to use closed or open population models. The obvious way to proceed is to fit the Jolly-Seber open population model first and use it to evaluate the closure assumption.

Table 1

Structure of the Population Under an Open Population Model*

Period	Pre-season List	Season Lists (*e.g.*, every two weeks)					
	0	1	2	3	. . .		k
Marked Population Sizes	M_0	M_1	M_2	M_3	. . .		M_k
Total Population Sizes	N_0	N_1	N_2	N_3	. . .		N_k

* Marked and Total Population Sizes are shown for the whole study.

3.2.1 Open Population Models

Under the Jolly-Seber model previously discussed in Section 2.3 the following parameters are identifiable (Table 2). Notice that it is possible to estimate the number of fishing boats in the fleet at each time in the season except the last (*i.e.*, $\hat{N}_k$ cannot be estimated). One advantage of applying the model in this fashion with a preseason list is that any concerns with the preseason list due to it being out of date are taken care of by the model allowing for additions and deletions before the season begins. One disadvantage of the Jolly-Seber Model is increased complexity. Now each time period has its own frame size and there are also survival and recruitment parameters to estimate. Sometimes these parameter estimates have poor precision unless sample sizes are large. Another disadvantage of the Jolly-Seber model is that it does require the assumption of equal catchability.

Table 2

Structure of the Jolly-Seber Open Population Model*

Period	Preseason	Season						
	0	1	2	3	.	.	k-1	k
Marked Population	$(M_0 \equiv 0)$	$\hat{M}_1$	$\hat{M}_2$	$\hat{M}_3$	.	.	$\hat{M}_{k-1}$	-
Total Population	-	$\hat{N}_1$	$\hat{N}_2$	$\hat{N}_3$	.	.	$\hat{N}_{k-1}$	-
Survival Rate	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$		.	$\hat{p}_{k-2}$		-
Recruitment No.		$\hat{B}_1$	$\hat{B}_2$		.	$\hat{B}_{k-2}$		

* Identifiable parameter estimators are shown for Marked Population Sizes, Total Population Size, Survival Rate and Recruitment Number.

Another important question about the use of the Jolly-Seber model is what is called "temporary emigration." A fishing boat might leave the fishery for some periods and then return. The Jolly-Seber model makes the assumption that fishing boats which leave do not return. This issue needs further investigation. Use of the robust design (*i.e.*, combination closed and open models) allows for temporary emigration. This would necessitate having two lists obtained close together in each period.

3.2.2 Closed Population Models

If the Jolly-Seber model estimates of "survival" and "recruitment" suggest population closure (*i.e.*, N constant) then the general closed population models reviewed in Section 2.2 could be applied. The advantages are increased precision of $\hat{N}$ due to the use of more lists and increased robustness of $\hat{N}$ to unequal catchability. The disadvantage is primarily an increase in complexity.

4. DISCUSSION

4.1 Methods of Dealing with Incomplete List Frames

(i) Complete the List Frame

The advantage is that the survey researcher has a complete frame and does not have to generalize results for an estimated frame size. The disadvantage is the cost and possible impracticality of completing the list frame.

(ii) Use an Area Frame

The advantage is that one only has to enumerate the establishments in the areas to be sampled. The disadvantage is possible inefficiency if businesses are sparse in each large area.

(iii) Using List and Area Frame (Multi-Frame Approach)

The advantages are obviously increased precision and having all establishments covered. The disadvantage could be expense and impracticality.

(iv) Use of Capture-Recapture to Estimate List Frame Size

The advantage is having a practical method of lower expense than the first three approaches listed above. The disadvantages are potential bias if the assumptions of the capture-recapture method are violated and having to include variation due to frame size estimation in variance estimates of population total estimates.

4.2 Capture-Recapture Estimation of Frame Size

In this section we consider model assumptions, precision of estimates, estimation of population totals and the special problems in more complex sampling designs when the capture-recapture approach to frame size estimation is used.

Model Assumptions

(i) Closure

Can the frame size be considered constant so that the closed population models be used? This will depend on whether the survey is just a snapshot at a single time point or whether a series of surveys over time are required. It will also depend on how quickly establishments go out of business and how quickly new ones arise. We suspect there will be the need for use of closed and open population models depending on the establishments being studied.

There is also the question of temporary emigration where establishments go out of the frame and then come back in again. This was considered a potential problem in the fishing boat example because boats could go inactive and then become active again. This may also be a problem in some other establishment surveys if establishments go in and out of business frequently and keep the same name when they come back into business.

(ii) "Unequal Catchability" and Independence of Lists

As we discussed earlier ideally the lists used should be independent so that the estimates of frame size are unbiased. In practice it may not be easy to find two or more independent lists.

(iii) Mark Loss-Unique Identification of Establishment

Establishment names need to be unique and unmistakable or matches on different lists may be missed or mistaken. This was a problem in the fishing boat example in earlier years. We suspect this will not be such a big problem in most establishment surveys.

Precision of Estimates

The lists used need to be of sufficient size that the precision of the frame size estimate ($\hat{N}$) is adequate. Seber (1982, p. 96) discusses the Lincoln-Petersen estimate in detail and presents graphics of sample sizes required for various levels of precision. Pollock *et al.* (1990) presents sample size information for the open population models.

Estimation of Population Totals

Once the estimate of frame size is obtained then that estimate will often be combined with a sample mean to obtain an estimate of a population total ($\hat{Y} = \hat{N}\bar{y}$). The estimate of population total is subject to possible bias and additional variance because $\hat{N}$ is estimated. The estimate may also be biased because $\bar{y}$ is not based on a random sample of the complete frame.

More Complex Sampling Designs

In this paper we have emphasized estimation of frame size in simple random sampling using the capture-recapture method. Further questions arise if more complex sampling designs are used. For example in stratified designs the question would arise of whether to estimate frame size in each stratum separately or to estimate the total frame size and then apportion it to the strata assuming equal probabilities of different strata on the incomplete lists. There is also the more complex question of how to estimate frame size in multi-stage sampling designs. This is obviously an area that needs future research.

ACKNOWLEDGEMENTS

The authors wish to thank the editor, associate editor and two anonymous referees for helpful comments which have greatly improved the paper.

REFERENCES

CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with application to zoological census. *University of California Publication in Statistics*, 1, 131-160.

COCHRAN, W.G. (1978). *Sampling Techniques* (Third Edition). New York: John Wiley and Sons.

COWAN, C.D., BREAKEY, W.R., and FISCHER, P.J. (1986). The methodology of counting the homeless. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-175.

DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.

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

GOODMAN, L.A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.

GREENE, M.A. (1983). Estimating the size of the criminal population using an open population approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-13.

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

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

JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52, 225-247.

KENDALL, W.L. (1992). Robust Design in Capture-Recapture Sampling: Modelling Approaches and Estimation Methods. Unpublished PhD. dissertation, North Carolina State University, Biomathematics Program.

MENKINS, G.E., and ANDERSON, S.H. (1988). Estimation of small mammal population size. *Ecology*, 69, 1952-1959.

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, 1-125.

POLLOCK, K.H. (1974). The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments. Unpublished Ph.D. dissertation, Cornell University, Biometrics Unit.

POLLOCK, K.H. (1982). A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46, 752-757.

POLLOCK, K.H., NICHOLS, J.D., HINES, J.E., and BROWNIE, C. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1-97.

POLLOCK, K.H., and OTTO, M.C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics*, 39, 1035-1049.

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

SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.

SEBER, G.A.F. (1965). A note on the multiple recapture census. *Biometrika*, 52, 249-259.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters* (Second Edition). New York: MacMillan.

SUDMAN, S., SIRKEN, M.G., and COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-995.

VOGEL, F.A., and KOTT, P. (1993). Multiple frame establishment surveys. *Proceedings, International Conference on Establishment Surveys*.

WHITE, G.C., ANDERSON, D.R., BURNHAM, K.P., and OTIS, D.L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos, NM: Los Alamos Laboratory.

WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 79-93.

WITTES, J.T., COLTON, T., and SIDEL, V.W. (1974). Capture-recapture method for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 125–136
Statistics Canada

Questionnaire Design for Business Surveys

A.R. GOWER[1]

ABSTRACT

This paper provides an overview of important considerations that should be taken into account when developing and designing questionnaires for business surveys. These considerations include the determination of objectives and data requirements, consultation with data users and respondents, and methods for testing questionnaires. In developing and designing business survey questionnaires, focus groups and cognitive research methods help the researcher to identify potential sources of measurement error and to understand the response process that respondents go through in completing the questionnaires. Examples of focus groups and cognitive research undertaken by Statistics Canada are provided.

KEY WORDS: Questionnaire testing; Focus groups; Cognitive research.

1. INTRODUCTION

There are many types of business survey questionnaires. Typically, a business survey questionnaire collects information about a company's employees, its inventories, inputs, products, sales, and finances. It may also involve the collection of information related to market research or client satisfaction.

Business surveys are conducted by mail or administered by an interviewer in person or over the telephone. Follow-ups to mail surveys are often conducted by telephone. New data collection technologies for business surveys involve computer-assisted interviewing, fax machines, touchtone self-response, and the electronic transmission of data.

As in other types of surveys, questionnaires play a central role in the data collection process in a business survey. They have a major impact on data quality and on the image that a survey organization projects to its respondents.

The purpose of this paper is to provide an overview of questionnaire design for business surveys. The paper discusses important considerations such as the determination of objectives and data requirements, consultation with data users and respondents, the nature and concerns of business survey respondents, and methods for testing questionnaires.

In developing and designing business survey questionnaires, it is especially important to understand the response process that respondents go through in completing the questionnaires. Therefore, this paper emphasizes the effectiveness of using focus groups and cognitive research techniques to develop and test business survey questionnaires. Examples of focus groups and cognitive research that have been carried out by the Questionnaire Design Resource Centre of Statistics Canada are provided.

2. BUSINESS SURVEY QUESTIONNAIRES

A well-designed questionnaire in a business survey should collect data efficiently, with a minimum number of errors. Moreover, questionnaires should facilitate the coding and capture of data. They should minimize the amount of editing and imputation that is required. They should also lead to an overall reduction in the cost and time associated with data collection and processing (Statistics Canada 1994).

There are many considerations that apply to the development and design of business survey questionnaires. One key consideration is the nature of the respondent population. Business survey respondents answer in their role as employers or employees of a business. How a questionnaire is completed depends on the position and level of responsibility that the respondent holds in the business organization or company. Therefore, it is critical to identify the most appropriate person to provide the information in a business survey.

Response burden is a very real concern for business survey respondents. It depends on the number of questions that are asked, the time required to complete the questionnaire, and the effort that respondents put into searching or manipulating other data sources to provide the information in the format requested.

Businesses vary in size. Large businesses may have employees whose responsibilities include completing government and survey forms. In small businesses, respondents are often the owners or office managers who may not have as much time or flexibility in their schedules to complete the questionnaire.

Information provided by respondents in business surveys typically involves the use of records or other information systems. Questionnaires often contain technical or professional terminology associated with providing financial or administrative data.

[1] A.R. Gower, Questionnaire Design Resource Centre, Statistics Canada, Ottawa, Ontario, K1A 0T6.

Another consideration is the confidentiality and sensitivity of the information that the questionnaire is collecting. In many cases, businesses are concerned about providing confidential financial information that they do not want to reveal to competitors, governments or any other party. Therefore, assurances of confidentiality should be provided. All necessary arrangements should be made for the proper handling and custody of data in order that the confidentiality of information is ensured.

3. THE RESPONSE PROCESS IN BUSINESS SURVEYS

The model of the response process is well-known for household surveys. Answering these types of questions involves comprehension, retrieval, thinking/judging, and responding (Tourangeau 1984). Respondents must first understand the question. They then search their memories to retrieve the requested information. After retrieving the information, they think about what the correct answer to the question might be and how much of that answer they are willing to reveal. Only then do they give an answer to the question.

A corresponding response model for business surveys has also been developed (Edwards and Cantor 1991). Although the business survey model is similar to the household survey model, there are differences. The major difference is that business survey respondents must normally access one or more external sources of information such as financial or administrative records.

The ability of respondents to retrieve the requested information depends upon their familiarity with and understanding of the external source of information. They must also understand the relationship between the survey questions and the external data source. Multiple sources of information may add to the difficulty or complexity of this task. Further complexities may be introduced if the respondent has to consult another individual who can provide the requested information and who, in turn, may have to use one or more data sources (Gower and Nargundkar 1991).

4. DEVELOPMENT AND TESTING OF BUSINESS SURVEY QUESTIONNAIRES

There are several basic steps that are involved in developing and testing business survey questionnaires. These steps are discussed below.

4.1 Determination of the Objectives and Data Requirements

A document should be prepared that provides a clear and comprehensive statement of the survey objectives, data requirements, and the data analysis plan. This document is a necessary step that leads to the determination of the variables to be measured, the survey questions, and the response alternatives.

When designing the questionnaire, it is important to determine and understand the rationale for each question, how the information will be used, and whether the questions will be good measures of what is required.

4.2 Consultation with Clients, Data Users, Subject Matter Experts, and Respondents

In formulating objectives and data requirements, consultation should take place with clients and data users to fully understand their requirements and expectations. Subject matter experts should be contacted for advice and guidance.

If possible, the survey researcher should consult members of the survey population. This will help identify issues and concerns that are important to respondents, and may affect decisions regarding the content of the questionnaire. In addition, consultation with respondents will identify the language and terminology that respondents themselves use and will help clarify terminology, concepts and definitions.

4.3 Previous Questionnaires

Examining questions that were used in other surveys on the same or a similar topic provides a useful starting point in formulating the questions and response categories. In some situations (*e.g.*, for comparing data over time), the same questions may be used. The researcher should ensure that the questions are phrased so as to provide valid, consistent, and effective measures of the variables of interest.

4.4 The Use of Focus Groups in Developing Questionnaires

A *focus group* is an informal discussion of a selected topic involving participants who are chosen from the survey population. It provides insights into the attitudes, opinions, concerns, and experiences of the participants. A focus group is led by a moderator who is knowledgeable about group interviewing techniques and the purpose of the discussion.

Focus groups provide the opportunity to consult respondents, data users, and interviewers. In the early stages of developing a questionnaire, focus groups are used to develop the survey objectives and data requirements, to identify salient research issues, and to clarify definitions and concepts.

Focus groups are also useful in testing and evaluating questionnaires (see 4.6 below). They are used to evaluate respondents' understanding of the language and wording used in questions and instructions, and to evaluate alternative question wordings and formats.

Recruiting participants from businesses poses unique challenges for focus groups. Monetary incentives or honoraria that are usually offered to focus group participants (currently in the order of $30 to $50 each) may not be appropriate for business people. Assurances of confidentiality and emphasis on the importance of the survey and their participation in the study are more meaningful. Another type of incentive that may be offered is a donation to a non-profit organization of the participant's choice. Statistics Canada often gives focus group participants a copy of a publication that is of interest to them.

Focus groups vary in size from 6 to 12 persons. The optimum size is 7 or 8 persons for business participants, although smaller groups with 4 or 5 people (called mini focus groups or mini groups) are sometimes held. Because of difficulties in finding participants from businesses, focus groups should be conducted at a time that is convenient to the participants. For business people, focus groups are often held during working hours. Focus groups are audio-recorded, and are viewed by observers in an adjoining room behind a one-way mirror. Participants are fully informed that audio-recording is taking place and that they are being observed.

4.5 Considerations in Drafting the Questions

Many considerations go into writing the questions and developing the response categories. It is important to keep in mind the objectives and data requirements as well as how the information will be collected and processed. The questions must relate to the information needs. They must be addressed to the right people in the organization or company.

The method of data collection will determine how the questions and response categories will be formulated. The question wording must be clear, and they must be ordered in a logical sequence. The questions must be designed to be easily understood and accurately answered by respondents. Response categories and time reference periods should be compatible with the business's record-keeping practices; however, this is often difficult to achieve.

The layout of the questionnaire should be attractive. The questionnaire should be *respondent-friendly* and, if administered by an interviewer over the telephone or in person, it should be *interviewer-friendly*.

The questionnaire should appear professional and "business-like". When designing the questionnaire, it should be kept in mind that businesses are asked to complete many forms and questionnaires. Completing them is not a priority. Research conducted by Statistics Canada's Questionnaire Design Resource Centre has shown that typical reactions from businesses to questionnaires are:

- "I complete the shortest form first."
- "Is completion mandatory?"
- "Is there a return deadline?"

In one Statistics Canada study (Gower and Zylstra 1990), a respondent commented that if the answer to these last two questions is "no," then "I put [the questionnaire] in my *maybe I'll get to it someday* basket!"

Respondents frequently question the value of information to themselves and to other users. Some like to receive feedback about the survey. Therefore:

- Explain why it is important to complete the questionnaire.
- Ensure that the value of providing information is made clear to respondents.
- Explain how the survey data will be used.
- Explain how respondents can access the data.

The instructions that go with the questionnaire also require attention. Research carried out by the Questionnaire Design Resource Centre has repeatedly shown that respondents read only what they *think* is necessary to read. They read the boldface print first, and then decide whether they should read further. Respondents rarely read the instructions, and usually proceed directly to the questions. They refer to the instructions only when they *think* they need help. As a result, respondents may miss important instructions and definitions. Errors in reporting are often due to a lack of clear instructions and due to respondents not reading them or not understanding them (*e.g.*, what to include or exclude). Therefore:

- Ensure that instructions are short and clear.
- Tell the respondent where to find the instructions.
- Provide definitions at the beginning of the questionnaire or in specific questions as required.
- Use **boldface print** or underlining to emphasize important items such as the reference or reporting period.
- Specify "include" or "exclude" in the questions and items themselves (not in separate instructions).

Other considerations that should be taken into account in designing business survey questionnaires include:

- Consistency of terminology, questions and response categories with standard concepts and definitions.
- Nature of the respondent population such as record-keeping practices and language ability.
- Availability of the data.
- Response burden.
- Complexity of the data to be collected.
- Comparability of results with other surveys.
- Data reliability.
- Nonresponse.

The design of the questionnaire should also take into account any administrative requirements of the survey organization. For example, Statistics Canada's policy on informing survey respondents (Statistics Canada 1986) requires that key information be explained to respondents. They must be informed about the main purpose(s) of the

survey, the major intended uses of the data, the requirement to respond (compulsory or voluntary), confidentiality protection, and any joint collection or data sharing agreements. At Statistics Canada there are also other administrative or legal requirements. For example, the Official Languages Act of Canada requires that questionnaires be made available to respondents in both official languages (*i.e.*, English and French).

4.6 The Use of Cognitive Methods in Testing Questionnaires

Questionnaire testing is essential to developing effective questionnaires that collect useful and accurate data. Cognitive research methods, sometimes referred to as qualitative testing, are especially useful in testing questionnaires.

Cognitive methods provide the means to examine respondents' thought processes as they answer the survey questions. They are used to ascertain whether or not respondents understand what questions mean and thus help assess the validity of questions and identify potential sources of measurement error. Cognitive methods also provide the opportunity to evaluate the questionnaire from the respondent's point of view. They focus on issues such as comprehension and reactions to the form. This brings the respondent's perspective directly into the questionnaire design process. The use of cognitive methods leads to the design of respondent-friendly questionnaires that can be completed easily and accurately.

In business surveys, cognitive methods are used to investigate the relationship between the respondent and the external information source. They are also used to study the influence that this data source has on the response process. These methods provide the means to assess the compatibility of question wording, time reference periods, and response categories with the business's record-keeping practices.

Cognitive testing methods (Gower 1993) include:

- *In-depth interviews*: The technique involves one-on-one interviews (sometimes called retrospective think-aloud interviews). For a mail questionnaire, respondents first complete the questionnaire as they normally would. An interviewer observes the process, noting the sequence in which the questions are answered, reference made to instructions, and the types of records or other persons consulted. The interviewer also notes the time required to complete sections, and corrections or changes made to responses.

  The interviewer then conducts the in-depth interview and obtains information about the respondent's experiences and impressions in completing the form. The follow-up discussion typically involves a question-by-question review of the questionnaire with the respondent to discuss any problems or difficulties that were encountered while completing the form. The interviewer probes to see how the terms and concepts were interpreted by the respondents, how and why they chose the responses, and how information was recalled.

  For an interviewer-administered questionnaire, the questions are first asked by an interviewer either in person or by telephone. The in-depth follow-up discussion takes place following this first interview.

- *Concurrent think-aloud interviews*: These are also conducted one-on-one. The respondent is asked to "think aloud" while answering the questions, commenting on each question and explaining how the final response was chosen. The observer may probe the responses to get more information about a particular statement or to clarify the process through which a response was chosen.

  The success of the concurrent think-aloud interview technique depends on the respondent's ability and willingness to articulate and express thoughts aloud. The observer may sometimes have to help the respondent in this task by gentle prompts such as: "what question are you answering now?", "what are you thinking now?", "please explain how you chose the answer", or other probes to clarify the respondent's thoughts. When a respondent is reluctant to verbalize thoughts, the observer may decide that the better approach is to handle the interview as an in-depth interview and proceed accordingly.

  Think-aloud interviews are very useful in obtaining respondents' reactions to questionnaires. They are especially helpful in identifying areas of the questionnaire where respondents have difficulty. They also help the researcher understand the process through which the questionnaire is completed.

- *Focus groups*: As described in 4.4, focus groups are used to evaluate respondents' understanding of the language and wording used in questions and instructions. The questionnaire is usually administered before the focus group session, in person, over the telephone or on a self-completion basis.

  During the focus group session, the moderator reviews the questionnaire with the participants and discusses any problems or difficulties that they may have encountered when completing the form. Focus groups stimulate and encourage thoughtful analysis of the questionnaire during group discussions of individual participants' comments. They are especially useful in providing suggestions and recommendations for improvements.

- *Paraphrasing*: Paraphrasing is used in one-on-one interviews and focus groups. Respondents are asked to repeat the question in their own words, or to explain the meaning of terms and concepts that are used in the survey questions and instructions.

Paraphrasing helps determine whether respondents read and understand the instructions and questions correctly. Paraphrasing is especially helpful in identifying question wording that is too complex or confusing. It also identifies situations where respondents do not comprehend all the important components of the question (*e.g.*, the reference period).

4.7 Pretesting

Pretesting is a fundamental step in developing a questionnaire. It usually involves a small number of field interviews that are carried out to identify problems with a questionnaire. The entire questionnaire or only a portion of it may be tested.

Pretests are useful for discovering poor question wording or ordering, errors in questionnaire layout or instructions, and problems caused by the respondent's inability or unwillingness to answer the questions. Pretests are also used to suggest additional response categories that can be precoded on the questionnaire. Pretests provide a preliminary indication of the interview length and refusal problems.

The pretest sample can range in size from 20 to 100 or more respondents. If the main purpose of the pretest is to discover wording or sequencing problems, only a small number of interviews may be required. More interviews (50 to 100) are necessary to determine pre-coded answer categories for open-ended responses. Respondents for pretests are usually selected purposively rather than randomly.

The questionnaire for a pretest should be administered in the same way as planned for the main survey (*e.g.*, interviewer-administered in person or by telephone). A pretest of a mail questionnaire is more effective if interviewers are used. Interviewers can be used to deliver the questionnaire and, later, to discuss any problems. The questionnaire designers should observe as many pretest interviews as possible.

Pretesting is not as effective as cognitive methods in evaluating respondents' understanding and the difficulty of the response task. Pretesting only indicates whether there is a problem. Without further investigation, it does not identify why there is a problem nor how it can be corrected.

Debriefing sessions with interviewers often occur in conjunction with a pretest. Interviewers involved in a pretest can identify important problem areas where the questionnaire can be improved. When existing questionnaires are redesigned, it is useful to consult interviewers to get their input into the redesign process. Interviewers have excellent insights into the logistics of administering the questionnaire and how it affects respondent cooperation.

Behavioral coding also can be conducted at the time of pretesting. The interview is audio-recorded, following which the interviewer and respondent behaviours during the interviewer-respondent interaction are coded and analyzed. Behavioral coding provides a systematic and objective means of examining the effectiveness of the questionnaire. It also helps to identify problem areas such as an interviewer failing to read the question as worded or a respondent asking for clarification of the question or response task.

4.8 Formal Testing Methods

Formal testing methods are quantitative in nature. They are designed to provide a statistical evaluation of how the questionnaire performs. Pilot studies and split sample testing are two commonly used types of formal testing methods. These methods are more suitable for large scale and continuing surveys because of the significant cost involved in implementing them and analyzing the results.

A *pilot study* is conducted to observe how all the survey operations, including the administration of the questionnaire, work together in practice. A pilot study is a "dress rehearsal". It duplicates the final survey design on a small scale from beginning to end, including data processing and analysis. It allows the survey researcher to see how well the questionnaire performs in relation to all other parts of the survey. There are some problems that can only be identified when all phases of the survey are tested together. For example, typographical errors and problems with question wording or concepts that need further clarification may be identified during interviewer training. The data processing phase may reveal keying problems with the precoded item numbers and/or answer categories (DeMaio 1983).

Normally, the questionnaire should be thoroughly pretested before a pilot study takes place. A pilot study is usually not the time to try out new questions or approaches. If previous testing has been carried out, it is unlikely that the pilot study will result in major changes to the questionnaire. The pilot study, however, does provide the opportunity to fine-tune the questionnaire before its use in the main survey (DeMaio 1983).

Split sample testing is conducted to determine the "best" of two or more alternative versions of the questionnaire. Split sample testing is also referred to as a "split ballot" or "split panel" experiment. It involves an experimental design that is incorporated into the data collection process. A split sample test can be designed to investigate issues such as question wording, question sequencing, the location of sensitive items, and data collection procedures. In a simple split sample design, half of the sample is selected at random and might receive one experimental treatment and half, the other. In a test that involves two experimental treatments, a 2×2 factorial design might be used with each of the two treatments in each experiment being tested on half of the sample (DeMaio 1983).

A split sample design can also be used in continuing surveys that assess trends over time and compare results across surveys. In these types of surveys, there often is a concern that any change in the questionnaire or procedures may affect other data items besides the items being added or revised. In these cases, a split sample design may be used with a random sample of the respondents receiving the "old" questionnaire and the rest, the "new" questionnaire. Comparisons with earlier data can still be made by using the old questionnaire for most or part of the sample (DeMaio 1983).

4.9 Review and Revision of the Questionnaire

The questionnaire should be reviewed by someone outside the project team. Reviewers could include subject matter experts or persons who have experience in designing questionnaires. A review can take place at any or all stages of the questionnaire development process, causing revisions in the questions and response categories.

Questionnaire design is an iterative process. Throughout the whole process of questionnaire development, revision and testing, changes will be made continually to improve the questionnaire. Objectives and information requirements are stated, evaluated and decided upon, data users and respondents are consulted, proposed questions are drafted and tested, questions are reviewed and revised, until a final questionnaire is developed.

5. APPLICATION OF FOCUS GROUPS AND COGNITIVE RESEARCH METHODS TO TEST BUSINESS SURVEY QUESTIONNAIRES

Statistics Canada has found that focus groups and cognitive research methods are very useful in developing and testing business survey questionnaires. These methods provide the opportunity to understand the cognitive processes involved in formulating responses to survey questions. They bring the respondent's perspective directly into the questionnaire design process and lead to the design of respondent-friendly questionnaires (Gower and Nargundkar 1991).

Statistics Canada's applications of focus groups and cognitive research methods for business surveys include the developing and testing of questionnaires for the following surveys:

- Survey of Employment, Payrolls and Hours (Bureau 1991; Goss, Gilroy and Associates Ltd. 1989; Goss, Gilroy and Associates Ltd. 1990).
- Census of the Construction Industry (Gower and Zylstra 1990; Price Waterhouse Management Consultants 1990).
- Wholesale and Retail Trades Survey (Noonan 1992).
- National Training Survey (Kennedy and de Groh Consultants 1992; D.R. Harley Consultants Limited 1993).

These studies involved the application of one or more of the following methods: focus groups, in-depth interviews, concurrent think-aloud interviews, and paraphrasing. All studies were carried out under the coordination and general direction of Statistics Canada's Questionnaire Design Resource Centre (Gower 1991).

Each of the studies has demonstrated the importance of and benefits to be gained from consulting with members of the target population before developing and finalizing the questionnaire. The studies have provided valuable insights into the response process and have identified various factors that contribute to measurement errors in business surveys. These factors include the respondents' perceived value of the information, their perception of response burden, the compatibility of questions with their record-keeping practices, the placement and use of instructions, the availability of data, and the complexity of the response task (Gower and Zylstra 1990).

Highlights from two of the studies, the Census of the Construction Industry and the National Training Survey, are discussed below.

5.1 Census of the Construction Industry

The annual Census of the Construction Industry was designed to provide comprehensive statistics on the construction industry in Canada. The target population consisted of establishments whose main revenue was derived from construction activity. There were two separate questionnaires for (a) General Contractors and Developers and (b) Trade Contractors and Sub-Contractors. The questionnaires, which were mailed to respondents, collected data on revenues and costs, labour data, and output distributions.

The questionnaires used in 1988 for the Census of the Construction Industry were redesigned for the 1989 survey. The main objectives of the revision were to reduce the content and response burden and to respond to the need for major improvements to the existing questionnaires.

A pretest of the revised questionnaires took place to obtain the reactions of contractors (Statistics Canada 1989). The pretest indicated that the revised forms were well received and understood by respondents. Some areas for further improvement such as changes to question wording and the clarification of certain instructions were identified.

To learn more about how respondents would view the revised questionnaires and to ensure that response rates and data quality would be maximized, further testing of the questionnaires using focus groups and cognitive methods was carried out in early 1990. This phase of testing was designed to obtain in-depth information on the following issues:

- How respondents felt about the questionnaires.
- The process that respondents went through to provide the information.
- The layout, presentation, and readability of the questionnaires.
- The extent to which respondents read and understood instructions and questions.
- Problems encountered by respondents while completing the questionnaires.
- Whether instructions and definitions were necessary, understandable, and useful.
- The accuracy of information provided by respondents.
- The use of estimates by respondents and their accuracy.
- The types of records from which information was obtained.
- The compatibility of the questions and response categories with respondents' record-keeping practices.
- Response burden in terms of time and effort.

The scope of the research included both the General Contractors and Developers questionnaire and the Trade Contractors and Sub-contractors questionnaire. Approximately 50 construction firms participated in the study. They were chosen to represent the types of respondents who completed the Census of the Construction Industry questionnaires. Twenty-five in-depth interviews, 16 concurrent think-aloud interviews, and 2 focus groups were conducted in Ottawa, Montréal and Toronto. All one-on-one interviews took place at the respondent's place of business.

A very interesting finding from the study was that there were two distinct groups of respondents. The first group of respondents included the president or vice-president of a company, who often had to consult other individuals to complete certain questions. It took these participants 35 to 45 minutes to complete the questionnaire. They were more likely to make estimates based on their familiarity with the company and were less concerned about accounting for differences between the questionnaire and the source of information used to complete the form.

On the other hand, respondents such as office managers, accountants and comptrollers took 75 to 90 minutes to complete the questionnaire. These respondents were much more concerned with detail and providing accurate answers. They were more likely to use multiple sources of information and to make calculations in answering the survey questions (Gower and Zylstra 1990; Gower and Nargundkar 1991).

Many respondents indicated that completing the questionnaire was not a priority. They viewed the survey as only one of the many forms and questionnaires that they had to complete each year. Many participants indicated that they often waited for the follow-up telephone call, and some even preferred, to answer the questionnaire over the telephone. They said that, over the telephone, they could make estimates "off the tops of their heads" instead of carefully completing the form, and this required much less time and effort on their part.

The response burden was more perceived than real. Upon completing the questionnaire, many respondents remarked that it took surprisingly less time and was easier to complete than they had anticipated.

A common theme that emerged during the interviews and focus groups was the perceived value of the information being collected. Respondents wanted to know the purpose of completing the questionnaire and often questioned the value of the information to themselves and to other users of the information. Therefore, a major finding of the research was that the value of providing the information must be made clear to respondents. They wanted to know how the survey results were going to be used. They were also interested in learning how they could access the data.

Overall, the questionnaires were very well received by respondents. They appreciated the "business-like" appearance and approach of the questionnaires. Many were familiar with completing previous questionnaires for the Census of the Construction Industry. They felt that the redesigned forms were an improvement over the previous versions because they seemed shorter and less complicated. This was positive feedback and reassurance for the survey managers who designed the new questionnaires (Gower and Zylstra 1990; Price Waterhouse Management Consultants 1990).

The study identified many specific findings about how the questionnaires could be improved and made more "respondent-friendly". While the pretest provided valuable feedback about response rates and the completeness of reporting, the focus groups and cognitive research added significantly to these findings by providing in-depth, first-hand information about *how* and *why* respondents reacted to the questions as well as about *how* and *why* responses were chosen.

Figures 1 and 2 illustrate a few of the specific findings and how the questionnaire was improved based on these findings (Gower 1993). Figure 1 shows parts of Sections 2 and 4 of the 1988 version of the questionnaire for General Contractors and Developers, *before testing*. Figure 2 shows the corresponding parts of the final version of this questionnaire, *after testing*.

Section 2 – Statement of Income

On the final version of the questionnaire (Figure 2):

- A statement is provided at the beginning of Section 2, telling respondents that they could include their company's Financial Statements. On the version of the form (Figure 1) that was tested, many respondents missed this instruction because it appeared on a separate page of instructions.

Figure 1 (*before testing*): 1988 Census of the Construction Industry (General Contractors and Developers), Statistics Canada

SECTION 2. STATEMENT OF INCOME

Dollars (Omit cents)

REVENUE

2.1 Revenue from construction contracts 101

2.2 Other operating revenue, please specify:

	Type		Value
102		103	$
104		105	
106		107	
108		109	
		Total	110

2.3 **Total gross operating revenue** (sum of items 2.1 and 2.2) 111

2.4 Accounting method is: 1 ☐ completed contract
2 ☐ percentage of completion

DIRECT COST

2.5 Work in progress, opening (add, if required for direct cost calculation) 112

If direct cost detail is not available, please report percentages of total (item 2.15, sum should equal 100).

		Percentage	or	
2.6	Sub-contracts			113
2.7	Materials and supplies used (adjusted for change in inventory)			114
2.8	Wages paid to hourly-rated employees (gross, before deductions for income tax, pension plans, insurance, etc.)			115
2.9	Direct salaries paid to site supervisors, etc. (gross, before deductions for income tax, pension plans, insurance, etc.)			116
2.10	Employee benefits (employer contributions not included in 2.8 and 2.9, such as pension plans, insurance, etc.)			117
2.11	Land			118
	Cost includes (please check): 1 ☐ undeveloped land; 2 ☐ services, carrying charges, etc.; 3 ☐ serviced lots			
2.12	Repair and maintenance of machinery and equipment			119
2.13	Equipment rental (without operator)			120
2.14	Other direct cost			121
2.15	**Total direct cost (sum of items 2.6 to 2.14)**	100		122
2.16	Work in progress, closing (deduct if required for direct cost calculation)			123
2.17	**Total direct cost charged to contracts (item 2.5 plus 2.15 minus 2.16)**			124

SECTION 4. LABOUR FORCE

4.1 For wages paid to your hourly paid labour force, reported in item 2.8, please report hours worked:

201 ☐ hrs. or average hourly rate: $ 202 ☐ / hour

N.B.: Reported figure should be hours worked, i.e. one hour overtime paid at time and a half should be counted as one hour.

4.2 For direct salaries paid, reported in item 2.9 please provide average annual number of employees:

203 ☐ employees

4.3 For overhead salaries paid, reported in item 2.19 please provide average annual number of employees:

204 ☐ employees

Figure 2 (*after testing*): 1989 Survey of the Construction Industry (General Contractors and Developers), Statistics Canada

SECTION 2. STATEMENT OF INCOME 201

Instead of completing this section, you may include your company's Financial Statements, together with your otherwise completed questionnaire. ***If financial statements are included, go directly to Section 3.***

		Dollars (Omit cents)
REVENUE		
2.1 Revenue from construction contracts	202	
2.2 Other operating revenue, such as sales of materials, land sales, project or construction management, rentals of equipment and buildings, snow removal, consulting engineering fees. *Please specify:*		
Description		
	203	207
	204	208
	205	209
	206	210
2.3 **Total gross operating revenue** *(sum of items 202 and 207-210)*		211
2.4 Please check accounting method used: 1 ☐ complete contract 2 ☐ percentage of completion	212	
DIRECT COSTS		
2.5 Work in progress, opening *(add, if required for direct cost calculation)*. Work in progress is defined as inventory of uncompleted and unbilled construction work done		213
Only if direct costs detail is not available, please estimate percentages of total direct costs *(item 234, sum should equal 100)*	Percentage	**or**
2.6 Sub-contracts *(include equipment rental with operator)*	214	224
2.7 Equipment rental without operator	215	225
2.8 Materials and supplies used *(adjusted for change in inventory)*	216	226
2.9 Wages paid to any hourly-rated employees *(gross, before deductions for income tax, pension plans, insurance, etc.)*	217	227
2.10 Direct salaries charged to contract and paid to permanent staff, such as foremen, site supervisors, etc. *(gross, before deductions for income tax, pension plans, insurance, etc.)*	218	228
2.11 Employer portion of employee benefits, such as pension plans and insurance. *(Report **only** if employee benefits are not included in wages and direct salaries above)*	219	229
2.12 Cost of land included in sales	220	230
2.13 Repair and maintenance of machinery and equipment	221	231
2.14 Depreciation charged to contracts	222	232
2.15 Other direct costs *(any other direct costs not separately reported above, such as pre-constuction costs, site costs, fees, advertising, fuel, etc.)*	223	233
2.16 **Total direct cost** *(sum of items 224 to 233)*	**100**	234
2.17 Work in progress, closing *(deduct if required for direct cost calculation) For definition of work in progress see question 2.5 above*		235
2.18 **Total direct costs charged to contract** *(item 213 plus 234 minus 235)*		236

SECTION 4. LABOUR FORCE

4.1 Please report hours worked by your hourly paid labour force *(whose wages were reported in item 227)*:

N.B.: Reported figure should be hours worked, i.e. one hour overtime paid at time and a half should be counted as one hour. Figures for hours worked may be obtained from payroll records or Workers Compensation Board reports.

401 ____ hours

Only if hours worked are not available,
please report average (straight-time) hourly rate:

402 $ ____ / hour

4.2 Please report the average annual number of direct salaried employees *(whose salaries were reported in item 228)*:

403 ____ employees *Exclude owners and partners of unincorporated businesses*

4.3 Please report the average annual number of overhead salaried employees *(whose salaries were reported in item 237)*:

404 ____ employees *Exclude owners and partners of unincorporated businesses*

4.4 Number of professional engineers included in item 404:

405 ____ engineers

- Reference is made to line numbers (*e.g.*, 202 and 207-210) instead of item numbers (*e.g.*, 2.1 and 2.2). Although the line numbers are actually data code numbers, respondents viewed them as line numbers because they appeared similar to the common and well-known use of line numbers on the Canadian Income Tax forms.
- Important information such as definitions and what to include are provided in the items themselves instead of on the Instructions page.
- Respondents are only required to report estimated percentages if detail about direct costs is not available. This choice has been made clearer by printing "**or**" in large and bold print.

Note that, in completing Section 2, respondents consulted the following types of records: financial statements, on-line accounting systems, progress or work-on-hand billings, project reports, general ledgers, working papers, and audit statements.

Section 4 - Labour Force

On the final version of the questionnaire (Figure 2):

- Question 4.1 includes information that "hours worked" may be obtained from "payroll records or Workers' Compensation Board reports". During the think-aloud interviews, respondents noted that they consulted these types of records for the information.
- Clarification is provided that "average hourly rate" is to be reported "only if hours worked are not available".
- Important information and instructions are included in the question items. For example, during testing, most respondents did not exclude owners and partners in reporting the numbers of employees in items 4.2 and 4.3 (even though this was specified on the Instructions page).

5.2 National Training Survey (NTS)

Two separate research studies, each involving the application of focus groups and cognitive research methods, have been used during the development and testing of the questionnaire for the National Training Survey (NTS).

The purpose of the NTS is to collect information on employee training and development in the private business sector. Respondents are asked to provide data on the type and volume of training, the number of trainees and their occupational groupings, the characteristics of the businesses providing training to their employees, and the amount of money being spent on this activity. In large businesses, respondents are the persons involved in the human resource planning and training areas of their company, while in smaller businesses they are typically the owner or chief executive officer.

At an early stage in developing the questionnaire, focus groups and in-depth interviews were held with representatives from small, medium and large companies. These methods were used because Statistics Canada felt it was important to consult representatives of the business community to ensure that their interests and concerns about training were considered in the design of the NTS questionnaire.

The focus groups and interviews evaluated the clarity and appropriateness of terminology and concepts associated with the training of employees within a business establishment. The study investigated respondents' understanding of terms such as "formal training" and "informal training" as well as their ability to use these terms to categorize their training activities.

Findings from this early phase of testing illustrated the importance of consulting with respondents before finalizing the terminology and concepts used in questionnaires. The findings from the study provided the survey project team with important information and insights into how the survey questions should be worded and how response options should be categorized.

For example, a significant finding from the focus groups and in-depth interviews was that many companies did not use the terms "formal" or "informal" to describe training activities and did not see the advantage or need to differentiate between the two terms. Many also perceived that there was no clear distinction between the terms "formal" and "informal" that would enable easy categorization of training activities.

The study helped the survey designers understand how respondents interpret terms and concepts. Participants provided suggestions on the appropriate terminology for them. For example, although they had difficulties with the terms "formal" and "informal," participants were able to provide characteristics to define these terms. They described formal training as having "a formal structured curriculum or course outline with a beginning, middle and an end; that it has known objectives or clearly defined goals; that it has an evaluation component; ....[and] that [it] has a dollar cost." On the other hand, most participants perceived "informal training" to be on-the-job training having no structure, often involving learning by observing. "Lack of evaluation" was another characteristic often suggested to define informal training.

Another interesting finding was that many participants made a distinction between "training" and "developmental or educational activities". The term "training" was not seen to cover all the activities that employers provide to support employee development. Some participants viewed "training" as job-specific and related to job productivity, and "development" as related to increasing the knowledge base of the individual (Kennedy and de Groh 1992).

After the draft NTS questionnaire was developed, it was tested using focus groups and concurrent think-aloud interviews. Representatives of a variety of businesses as well as a mixture of small, medium and large firms participated in the study. The study examined the following issues:

- The most appropriate person within a business to respond to the survey.
- How best to reach respondents.
- The process that respondents went through to provide the information.
- The way in which respondents understood the questions and instructions.
- Respondents' reaction to vocabulary and the groupings and classifications of occupations in the survey.
- Whether the information sought in the survey was readily available.
- The types of records from which information was obtained.
- The compatibility of the questions and response categories with respondents' record-keeping practices.
- Whether the reference periods requested in the survey corresponded to the record-keeping practices of respondents.
- Response burden in terms of time and effort.

Seven focus groups and 26 interviews were conducted in Ottawa, Toronto, Montréal, and Vancouver. In the final report (D.R. Harley Consultants Limited 1993), the Contractor reported many findings and made several recommendations to improve the questionnaire.

As in other studies of business surveys, a major finding was that many participants questioned the purpose behind the survey. They wanted to know why the information was being collected and how the survey results were going to be used. A strong theme that emerged throughout the focus groups and interviews was that respondents wanted to know "What's in this for me?"

Some participants suggested that the data be aggregated nationally, provincially and by sector so that they could compare themselves to other companies in their areas of business and in their part of the country. As one respondent said, "I would want the data to be specific to our industry with the volume and type of training that's being provided .... It should allow us to compare ourselves to others in our sector – number of employees being trained and the percentage of payroll being spent on employee training."

Many small and medium-sized business respondents found the questionnaire too broad and the level of detail too complicated for them to answer. In their opinion, the questionnaire was designed for larger organizations. For example, many small businesses felt that they could not fit themselves into the categories provided by the questionnaire. They felt that much of their training fell into the "unstructured" category, and that the questionnaire was not capturing this aspect of training. However, at the same time, there were other respondents from small and medium-size businesses who commented that the questionnaire was thorough and complete.

The larger businesses also had difficulty with the level of detail being requested by the survey. The major problem was that they keep training records by type of training that employees receive rather than by the occupational category of the people being trained.

Overall, a variety of record-keeping practices were observed. Some businesses keep excellent records on training, while others do not. Participants, who did not keep good records or whose records did not contain the requested information, found the questionnaire difficult to answer. Others, who had sophisticated records, could manipulate their data to fit the questionnaire. The one exception was the questions on training expenditure for which they found it difficult to provide detailed information. Global figures were more easily available, they said. Many businesses indicated that their training records were not centralized, thus making the questionnaire more difficult and requiring longer time to complete. They said that they would complete what they could, and then coordinate the completion of the rest of the questionnaire by forwarding it to many parts of their organization.

Although many participants were initially overwhelmed by the size and apparent complexity of the questionnaire, they found it easier to complete than expected. Many found that the thoroughness of the questionnaire actually made them remember many training activities that they would not ordinarily have reported on.

Most participants felt that the questionnaire should be shorter. But they also suggested adding a few more open-ended questions about future training. In terms of response burden, respondents (especially in medium-sized and large-size companies) found that the questions about training expenses, training hours, and the numbers of employees trained by occupational categories would require hours of work to compile.

Differences were found in the time it took respondents to complete the questionnaire. Small businesses took between 10 minutes and 1 hour to complete the questionnaire. Large businesses, on the other hand, estimated that it would take about 2 hours to complete the questionnaire (D.R. Harley Consultants Limited 1993).

6. CONCLUDING REMARKS

This paper has provided an overview of questionnaire design for business surveys. As the paper has pointed out, many considerations go into designing business survey questionnaires. They include the survey's objectives and data requirements as well as consultation with data users and respondents on the nature and concerns of the respondent population. Other considerations are response burden, the method of data collection, the availability of data, and the use of records, as well as the need for testing the questionnaires.

Specific design issues that should be taken into account include the instructions, the clarity and readability of the

questions, the logical sequencing of the questions, the compatibility of response categories and reference periods with respondents' record-keeping practices, and data processing requirements. The questionnaire should be respondent-friendly and interviewer-friendly.

To ensure the collection of accurate and useful data in business surveys, it is important to understand the response process that respondents go through in completing a questionnaire. Focus groups and cognitive research methods are very effective ways to study this response process and to test questionnaires. They provide the opportunity to consult directly with respondents and, thereby, to bring their ideas, concerns, and suggestions into the questionnaire design process.

Looking towards the future, research and experience should lead to improvements in the methods and approaches that are currently used to develop and test business survey questionnaires. An important area that requires more research and development is the relationship among the questionnaire, the respondent, and the external information source as well as the influence that this relationship has on the response process and the accuracy of reporting.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the work of the following consultants and contractors in undertaking this research: D.R. Harley Consultants Limited, Kennedy and de Groh Consultants, and Price Waterhouse Management Consultants. The views expressed are those of the author, and do not necessarily reflect those of Statistics Canada nor these contractors. The author also wishes to thank the referee for helpful comments.

REFERENCES

BUREAU, M. (1991). Experience with the use of cognitive methods in designing business survey questionnaires. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-717.

DEMAIO, T.J. (Ed.) (1983). *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10, Washington, DC: United States Office of Management and Budget.

EDWARDS, W.S., and CANTOR, D. (1991). Toward a response model in establishment surveys. In *Measurement Errors in Surveys*. (Eds. Paul P. Biemer *et al.*). New York: John Wiley and Sons, 211-233.

GOSS, GILROY and ASSOCIATES LTD. (1989). Qualitative Research to Evaluate the Questionnaire of the Survey of Employment, Payrolls and Hours (SEPH). Final report submitted to Statistics Canada.

GOSS, GILROY and ASSOCIATES LTD. (1990). Qualitative Research to Evaluate the Redesigned Survey Materials of the Survey of Employment, Payrolls and Hours (SEPH). Final report submitted to Statistics Canada.

GOWER, A.R. (1993). Questionnaire design for establishment surveys. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 950-956.

GOWER, A.R. (1991). The Questionnaire Design Resource Centre's Role in Questionnaire Research and Development at Statistics Canada. 48th Session of the International Statistical Institute. *Booklet*, Volume III, 58-59.

GOWER, A.R., and NARGUNDKAR, M.S. (1991). Cognitive Aspects of Questionnaire Design: Business Surveys versus Household Surveys. *Proceedings of the 1991 Annual Research Conference*. Washington, DC: United States Bureau of the Census, 299-312.

GOWER, A.R., and ZYLSTRA, P.D. (1990). The Use of Qualitative Methods in the Design of a Business Survey Questionnaire. Contributed Paper (unpublished). International Conference on Measurement Errors in Surveys, Tucson, Arizona.

D.R. HARLEY CONSULTANTS LIMITED (1993). Qualitative Testing of the Draft National Training Survey Questionnaire. Final report submitted to Statistics Canada.

KENNEDY and DE GROH CONSULTANTS (1992). Testing of Definitions for the National Training Survey. Final report submitted to Statistics Canada.

NOONAN, M. (1992). Final report on Personal Interviews with Potential Respondents for the Proposed Wholesale and Retail Trades Survey. Unpublished Report, Statistics Canada.

PRICE WATERHOUSE MANAGEMENT CONSULTANTS (1990). Qualitative Research Related to the Re-design of the Census of the Construction Industry Questionnaires. Final report submitted to Statistics Canada.

STATISTICS CANADA (1989). Construction Census Questionnaire Test. Unpublished Report, Construction Census Section, Industry Division.

STATISTICS CANADA (1994). Policy on the Development, Testing and Evaluation of Questionnaires.

STATISTICS CANADA (1986). Policy on Informing Survey Respondents.

TOURANGEAU, R. (1984). Cognitive sciences and survey methods. In *Cognitive Aspects of Survey Methodology: Building a Gap Between Disciplines*. (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, DC: National Academy Press, 73-100.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 137–147
Statistics Canada

Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse

E. RANCOURT, H. LEE and C.-E. SÄRNDAL[1]

ABSTRACT

Most surveys suffer from the problem of missing data caused by nonresponse. To deal with this problem, imputation is often used to create a "completed data set", that is, a data set composed of actual observations (for the respondents) and imputations (for the nonrespondents). Usually, imputation is carried out under the assumption of unconfounded response mechanism. When this assumption does not hold, a bias is introduced in the standard estimator of the population mean calculated from the completed data set. In this paper, we pursue the idea of using simple correction factors for the bias problem in the case that ratio imputation is used. The effectiveness of the correction factors is studied by Monte Carlo simulation using artificially generated data sets representing various super-populations, nonresponse rates, nonresponse mechanisms, and correlations between the variable of interest and the auxiliary variable. These correction factors are found to be effective especially when the population follows the model underlying ratio imputation. An option for estimating the variance of the corrected point estimates is also discussed.

KEY WORDS: Conditional bias; Monte Carlo simulation; Restoring estimator; Variance estimation.

1. INTRODUCTION

Occurrence of nonresponse is rather a norm than an exception in surveys. Missing data caused by nonresponse are often imputed to obtain a completed data set and the standard estimator is applied to the completed data set assuming that the underlying response mechanism is unconfounded. However, a point estimate obtained in such a way is biased when the response mechanism is confounded. The bias in this case could be very severe as pointed out in Lee, Rancourt and Särndal (1994). A response mechanism is unconfounded, according to Rubin (1987, p. 39), if it does not depend on the variable under study, otherwise it is confounded. (A formal definition suitable for this paper will be given in Section 2.)

In a Bayesian framework, a concept similar to that of an unconfounded response mechanism is termed ignorable. For bias caused by a nonignorable response mechanism, Rubin (1977, 1987) and Little and Rubin (1987) considered a method to correct the respondent mean using auxiliary variables. In this approach, a linear regression is assumed between the variable of interest y and a vector of auxiliary variables x. The regression coefficient vector for the nonrespondents is assumed to have a normal prior with mean equal to the regression coefficient vector for the respondents.

Assuming a logistic model for the response probability, Greenless, Reece and Zieschang (1982) proposed a method to deal with nonignorable nonresponse using maximum likelihood estimation. Further, a linear regression model is assumed for the relationship between y and x, a vector of auxiliary variables. The logistic model of the response probability includes y and z, a vector of other auxiliary variables. Assuming also that the error term of the regression is normally distributed, they obtain maximum likelihood estimates of the unknown parameters of the regression model and the logistic model. Finally, for a nonrespondent, an imputed value is calculated as the mean of the distribution of y conditional on the values of x and z for the nonrespondents, and the estimated parameters. Such a method may give good results when all the model assumptions are satisfied but is likely to be highly sensitive to the specifications of the two models. The adequacy of the response probability model is usually untestable. If data are available from an external source, however, then it may be possible to test the response probability model as Greenless *et al.* did in their application to the Current Population Survey data. This method is highly computer-intensive.

In the case of categorical data, a few methods have also been proposed to deal with the problem of nonignorable nonresponse. For instance, Baker and Laird (1988) try to model the response mechanism with the help of log-linear models. As well, causal modeling is discussed in Fay (1986, 1989).

Ratio imputation is often used at Statistics Canada, especially in repeated surveys. For instance, in the Monthly Survey of Manufacturing, a missing value of the current shipment is imputed by ratio imputation using previous month shipment as the auxiliary variable value. This simple method is very appealing to subject matter specialists because it reflects month-to-month movement.

[1] E. Rancourt and H. Lee, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; C.-E. Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec), Canada, H3C 3J7.

In this paper, we investigate the possibility of improving the estimator applied to data containing ratio imputation with the aid of simple correction factors. Therefore, we assume that imputation has already been performed, and try to correct the estimator. We focus our attention on the estimation of the mean. The use of simple correction factors would be very appealing to the user provided it works reasonably well. Such a procedure is also easy to implement without resorting to excessive computational efforts and it enables us to avoid explicit modeling of the nonresponse mechanism. However, our approach differs from Rubin's in that we use sample dependent correction factors rather than an *a priori* chosen constant.

In Section 2, we define several simple correction factors that meet our requirements. In Section 3, we propose a variance estimator that may be used in conjunction with the corrected point estimators. The properties of the corrected point estimators were examined by a Monte Carlo simulation reported in Sections 4 and 5. Section 6 presents some concluding remarks.

2. SIMPLE BIAS CORRECTION FACTORS

Let $U = \{1, \ldots, k, \ldots, N\}$ denote the index set of a finite population and let the population mean of the variable of interest y be denoted by $\bar{y}_U = (1/N)\sum_U y_k$. We assume that $y_k > 0$ for all $k \in U$. From U, a simple random sample s of size n is drawn without replacement (SRSWOR). The unbiased estimator that would be used with 100% response is the sample mean

$$\bar{y}_s = (1/n)\sum_s y_k. \tag{2.1}$$

Let r and o be the sets of the responding and nonresponding units, respectively, so that $s = r \cup o$. We denote the SRSWOR sampling plan by $p(\cdot)$ and the response mechanism given s by $q(\cdot \mid s)$. That is, $p(s)$ is the probability that the SRSWOR sample s is drawn, and $q(r \mid s)$ is the probability that the set r responds given the sample s. Let also m and l be the sizes of r and o, respectively. For simplicity, we assume that the probability of $m = 0$ is negligible. We assume that imputation is carried out with the aid of an auxiliary variable, x, whose value, x_k, is known and positive for all $k \in s$. If $k \in o$, the missing value y_k is imputed by $\hat{y}_k$. The completed data set is denoted as $\{y_{\cdot k} : k \in s\}$ where $y_{\cdot k} = y_k$ if $k \in r$ and $y_{\cdot k} = \hat{y}_k$ if $k \in o$.

In this paper, we examine ratio imputation. This often-used imputation method is based on a simple model. That is, if the value y_k is missing, it is imputed by $\hat{B}_r x_k$, where $\hat{B}_r = (\sum_r y_k)/(\sum_r x_k)$. The model denoted ξ, is stating that, for $k \in s$.

$$y_k = \beta x_k + \epsilon_k, \quad E_\xi(\epsilon_k \mid x_k) = 0, \quad V_\xi(\epsilon_k \mid x_k) = \sigma^2 x_k,$$
$$E_\xi(\epsilon_k \epsilon_l \mid x_k, x_l) = 0, \quad k \neq l. \tag{2.2}$$

Under this model, $\hat{B}_r x_k$ is the best linear unbiased predictor of the missing value y_k, based on the respondent data $\{(y_k, x_k) : k \in r\}$. The completed data set is then composed of the values

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}_r x_k, & \text{if } k \in o. \end{cases} \tag{2.3}$$

The customary procedure is to apply the estimator formula used for 100% response to the completed data set. This gives

$$\bar{y}_{\cdot s} = \frac{1}{n}\sum_s y_{\cdot k} = \frac{\bar{y}_r}{\bar{x}_r}\bar{x}_s = \bar{y}_{\text{raimp}}, \tag{2.4}$$

where $\bar{x}_s = (1/n)\sum_s x_k$, $\bar{y}_r = (1/m)\sum_r y_k$ and $\bar{x}_r = (1/m)\sum_r x_k$. Note that raimp stands for ratio imputed.

It now becomes necessary to address the question whether the imputation can restore the full response estimator, $\bar{y}_s$, in the sense that the imputation estimator $\bar{y}_{\cdot s}$ is equal to $\bar{y}_s$ in expectation given s. Unless this can be achieved, the ratio imputation will have introduced bias. To examine this question, we must consider the response mechanism. A response mechanism $q(\cdot \mid s)$ is said to be *unconfounded* for the purpose of this paper if it is of the form $q(r \mid s) = q(r \mid x_s)$, where $x_s = \{x_k : k \in s\}$ and the response probabilities satisfy $P(k \in r \mid s) > 0$ for all $k \in s$. That is, it may depend on s and on the associated x-values. If it depends also on the y-values, so that $q(r \mid s) = q(r \mid x_s, y_s)$, then is is called *confounded*. In these definitions, the response mechanism is conditional on the realized sample s. Slightly different definitions of "confounded" and "unconfounded" are given in Rubin (1987, p. 39) where they are unconditional.

An example of an unconfounded response mechanism is

$$q(r \mid s) = \prod_{k \in r}(1 - \Theta_k)\prod_{k \in s-r}\Theta_k,$$

where $\Theta_k = 1 - P(k \in r \mid s) = 1 - e^{-\gamma x_k}$ for some positive constant γ, is the nonresponse probability of unit k. By contrast, if $\Theta_k = 1 - e^{-\gamma y_k}$, then $q(r \mid s)$ is a confounded mechanism.

A particularly simple unconfounded mechanism is the uniform response mechanism defined by $q(r \mid s) = (1 - \Theta)^m \Theta^{n-m}$. Here, units respond according to independent and identical Bernoulli $(1 - \Theta)$ trials, where Θ is the nonresponse probability common to all units.

Whether an imputation estimator $\hat{\bar{y}}_U$ of $\bar{y}_U$, including $\bar{y}_{\text{raimp}}$ given by (2.4), is considered good depends in part on the assumptions made by the analyst about the response mechanism and in part on the relation between y and x. Several possible assumptions are discussed later in this section. For any given s, the goal is that, under specified realistic assumptions, the expectation of the difference

$\hat{\bar{y}}_U - \bar{y}_s$ should be close to zero. That is, under the given assumptions, the conditional bias of $\hat{\bar{y}}_U$, $C\text{-bias}(\hat{\bar{y}}_U) = E(\hat{\bar{y}}_U - \bar{y}_s \mid s)$, should be small. We call $\hat{\bar{y}}_U$ a *restoring estimator* of $\bar{y}_U$ if $C\text{-bias}(\hat{\bar{y}}_U) = 0$ or ≈ 0, that is, if $\hat{\bar{y}}_U$ is (approximately) equal to $\bar{y}_s$ in conditional expectation. It follows that if the C-bias is (approximately) zero for any s, then the unconditional bias over all sample realizations s is also (approximately) zero.

Different analysts make different assumptions. Let us consider some typical assumptions and ask the question: What restoring estimators do these assumptions allow?

Assumption I: The response mechanism is uniform.

Under Assumption I, $\bar{y}_{\text{raimp}}$ is a restoring estimator. To see this, note that

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = E_q(\bar{y}_{\text{raimp}} \mid s) - \bar{y}_s \approx 0,$$

because, given s, $\bar{y}_{\text{raimp}}$ is the classical ratio estimator of $\bar{y}_s$. Assumption I is unrealistic in most surveys. The response propensity is known to vary with observable characteristics such as size and industry (for business establishments), family size and type (for households), age, sex and income (for individuals). Under this unrealistic assumption, even a naive estimator such as the respondent mean, $\bar{y}_r = (1/m)\sum_r y_k$, is a restoring estimator:

$$C\text{-bias}(\bar{y}_r) = E_q(\bar{y}_r \mid s) - \bar{y}_s = 0.$$

However, if Assumption I holds, $\bar{y}_{\text{raimp}}$ is preferred to $\bar{y}_r$ because the ratio estimator feature leads to a smaller variance if the model ξ holds.

The analyst clearly needs to consider more realistic assumptions which allow the response probabilities to vary with background variables. The following assumption, composed of two parts, is of this kind.

Assumption II: (II-1): the response mechanism is unconfounded but otherwise arbitrary;
(II-2): the ratio model (2.2) holds.

Here (II-1) is a weaker and more realistic requirement on the response mechanism than the uniformity requirement in Assumption I. Under (II-1), the response mechanism can be of any form as long as it is unconfounded. However, Assumptions I and II are not directly comparable since II contains a model component, (II-2), which is lacking in I. Under Assumption II, $\bar{y}_{\text{raimp}}$ is a restoring estimator because

$$\begin{aligned} C\text{-bias}(\bar{y}_{\text{raimp}}) &= E_\xi\{E_q(\bar{y}_{\text{raimp}}) - \bar{y}_s \mid s\} \\ &= E_q E_\xi\left(\frac{\bar{y}_r}{\bar{x}_r}\bar{x}_s\right) - E_\xi(\bar{y}_s) \\ &= E_q(\beta\bar{x}_s) - \beta\bar{x}_s = 0. \end{aligned}$$

Note that changing the order of the expectations, $E_\xi E_q$ to $E_q E_\xi$, is allowed under Assumption II, because the response mechanism is then of the form $q(r \mid x_r)$, that is, it does not depend on the y-values. By contrast, the respondent mean $\bar{y}_r$ is not a restoring estimator because

$$C\text{-bias}(\bar{y}_r) = E_\xi\{E_q(\bar{y}_r) - \bar{y}_s \mid s\} = \beta\{E_q(\bar{x}_r \mid s) - \bar{x}_s\},$$

which is generally nonzero under Assumption II. We can, however, transform $\bar{y}_r$ into a restoring estimator by the use of a multiplicative correction factor. This leads to

$$\bar{y}_r\left\{1 + \left(1 - \frac{m}{n}\right)\left(\frac{\bar{x}_o}{\bar{x}_r} - 1\right)\right\}, \tag{2.5}$$

which is just another way of writing $\bar{y}_{\text{raimp}}$, as can easily be verified. In an example using the Bayesian approach, Little and Rubin (1987, p. 233) arrive at an estimator identical to the estimator (2.5).

Let us now consider confounded response mechanisms. They cause more difficult problems for finding a restoring estimator.

Assumption III: (III-1): the response mechanism is confounded but otherwise arbitrary;
(III-2): the ratio model (2.2) holds.

It is usually difficult, if not impossible, for the analyst to decide whether Assumption II or Assumption III is more appropriate. Examining the data will not be of much help if the only data available relate to the present point in time, as would typically be the case in a one-time survey. The assumption made (whether II or III) is then unverifiable. By contrast, if the analyst has experience with a regularly repeated survey, he or she may have legitimate reasons to believe, for example, that the nonresponse is a function of the variable of interest.

In some situations, the assumption of a confounded mechanism may be made on the following grounds. Suppose in a survey of personal finances that y, the variable under study is "savings" and that x, the auxiliary variable is "income", with values x_k known for the individuals $k \in s$. The nonresponse probability of respondent k is likely to be correlated with the savings figure y_k that he or she is asked to reveal as well as with the income figure x_k known from other sources. But since savings, not income, is the variable with which the respondent is directly confronted in the survey, the assumption that the nonresponse probability is a function of y_k may be more realistic than the assumption that it is a function of x_k. Hence a confounded mechanism may be more realistic to assume than an unconfounded mechanism.

Under Assumption III, neither $\bar{y}_r$ nor $\bar{y}_{\text{raimp}}$ are restoring estimators. The C-bias of $\bar{y}_{\text{raimp}}$ can be expressed as

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = \bar{x}_s E_\xi E_q \left(\frac{\sum_r \epsilon_k}{\sum_r x_k} \right),$$

where ϵ_k is defined by the model (2.2). This C-bias is generally nonzero and can be quite large when the nonresponse rate is high and the correlation is not so strong. However, the C-bias is hard to evaluate, since the exact form of the response mechanism is left unspecified. Note that changing the order of the expectations E_ξ and E_q is not permitted under Assumption III since $q(r \mid s)$ depends on the y-values. For example, a negative C-bias is likely to occur if the respondent residual total, $\sum_r \epsilon_k$ tends to be negative.

A confounded response mechanism (as in Assumption III), introduces bias in the slope estimator $\hat{B}_r = (\sum_r y_k)/(\sum_r x_k)$. Consequently, $\hat{B}_r x_k$ is a biased imputation for a missing value y_k. To improve the situation, suppose that a missing value y_k is imputed by $C\hat{B}_r x_k$ instead of $\hat{B}_r x_k$, where C is a quantity to be specified. Then the data after imputation are given by

$$y^c_k = \begin{cases} y_k, & \text{if } k \in r \\ C\hat{B}_r x_k, & \text{if } k \in o \end{cases} \qquad (2.6)$$

and denoting the sample mean of these data as $\bar{y}_{c\cdot s} = (1/n)\sum_s y^c_k$, we get the estimator

$$\bar{y}_{c\cdot s} = \bar{y}_r \left[1 + \left(1 - \frac{m}{n} \right) \left(C \frac{\bar{x}_o}{\bar{x}_r} - 1 \right) \right]. \qquad (2.7)$$

A simple correction of the type used in (2.6) was mentioned in Rubin (1986; 1987, p. 203) in the context of multiple imputation. Rubin views C as a fixed constant chosen by the user according to his or her prior knowledge. If such a choice happens to be well founded, the bias of (2.7) may be small.

Here, we shall examine choices of C that are adaptive, that is, they reflect the realized sample s and the realized response set r. Ideally, C should be such that the imputation will exactly restore the estimator $\bar{y}_s = (1/n)\sum_s y_k$ that would be used with 100% response. This C-value is determined by the equation

$$\bar{y}_s = \frac{1}{n}\sum_s y_k = \frac{1}{n}\sum_s y^c_k = \frac{1}{n}\left(\sum_r y_k + \sum_o C\hat{B}_r x_k \right).$$

A simple calculation shows that the optimal C-value is

$$C_{\text{opt}} = \frac{\hat{B}_o}{\hat{B}_r},$$

where $\hat{B}_o = \sum_o y_k / \sum_o x_k$ is the slope estimate if the model (2.2) could be fitted to nonrespondents. The imputed values would then be $\hat{y}_k = \hat{B}_o x_k$ for $k \in o$. Obviously, C_{opt} and $\hat{B}_o$ cannot be computed since they depend on missing y_k-values. For an unconfounded mechanism (as in Assumption II), we can expect $C_{\text{opt}} \approx 1$, given s, because

$$E_\xi E_q(C_{\text{opt}} \mid s) = E_q E_\xi \left(\frac{\hat{B}_o}{\hat{B}_r} \mid s \right) \approx 1.$$

But for a confounded mechanism (as in Assumption III), C_{opt} can be distinctly away from unity. Suppose that $C_{\text{opt}} > 1$. Note that $C_{\text{opt}} > 1$ if and only if $\sum_r e_{ks} < 0$ with $e_{ks} = y_k - \hat{B}_s x_k$, where $\hat{B}_s = (\sum_s y_k)/(\sum_s x_k)$ is the unknown slope estimate with 100% response. That is, $C_{\text{opt}} > 1$ implies that respondents' residuals e_{ks} are negative on the average. An illustration of this is shown in figure 1, where $n = 10$, $l = n - m = 5$, and all five respondents' residuals e_{ks} are negative.

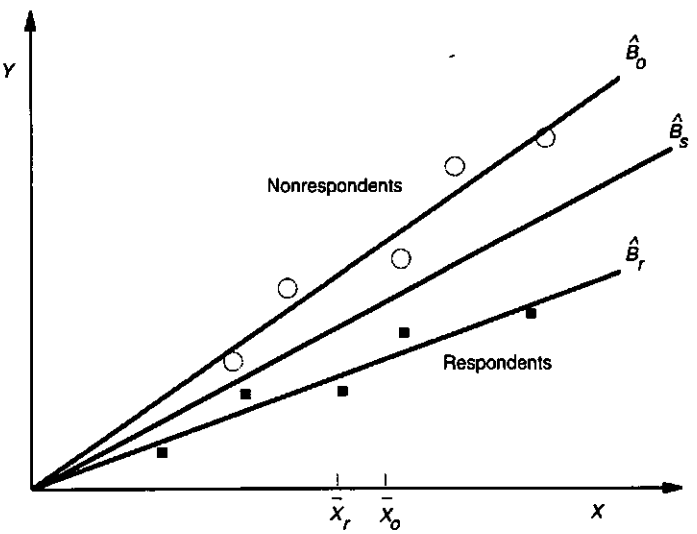

Figure 1. Example of data plot (y_k, x_k) for a confounded response mechanism.

Assuming that $C_{\text{opt}} > 1$, one approach for the analyst working under Assumption III is to choose a computable C likely to satisfy $C > 1$ and then use this C to construct the estimator (2.7). Factors C that will sometimes work in this manner are

$$c_1 = \frac{\bar{x}_o}{\bar{x}_r}, \quad c_2 = \frac{\bar{x}_o}{\bar{x}_s}, \quad c_3 = \frac{\bar{w}_o}{\bar{w}_r}, \quad c_4 = \frac{\bar{w}_o}{\bar{w}_s}. \qquad (2.8)$$

They are based on the logic that if the response mechanism is confounded in such a way that the nonresponse probability is a function of y (for example, $\Theta_k = 1 - e^{-\gamma y_k}$

with $\gamma > 0$), then both $C_{opt} > 1$, and $\bar{x}_o > \bar{x}_r$ are likely to occur, as Figure 1 illustrates. Conversely, if nonresponse is a decreasing function of y_k, then both $C_{opt} < 1$, and $\bar{x}_o < \bar{x}_r$ are likely to occur.

One important feature of such correction factors is that they can, but need not, be calculated during the imputation phase. For instance, if the usual ratio imputation $\hat{B}_r x_k$ was carried out at the imputation phase, it is then possible to calculate a suitable correction factor at the estimation phase without changing the originally imputed values.

Note that c_2 implies a somewhat milder correction than c_1: if $c_1 > 1$, we have $1 < c_2 < c_1$. The choices $C = c_3$ and $C = c_4$ are calculated on the ranks of the x-values, rather than on the x-values themselves, to dampen the effect of extreme x-values. More specifically, letting w_k be the rank of x_k in the data set $\{x_k : k \in s\}$, the w-means in c_3 and c_4 are $\bar{w}_s = (1/n)\sum_s w_k$, $\bar{w}_r = (1/m)\sum_r w_k$ and $\bar{w}_o = (1/l)\sum_o w_k$. The four estimators obtained by letting $C = c_i$ in (2.7) according to (2.8) will be denoted as $\bar{y}_{c_i \cdot s}$, $i = 1, \ldots, 4$. In particular, we have

$$\bar{y}_{c1 \cdot s} = \bar{y}_r\left[1 + \left(1 - \frac{m}{n}\right)\left\{\left(\frac{\bar{x}_o}{\bar{x}_r}\right)^2 - 1\right\}\right], \quad (2.9)$$

and

$$\bar{y}_{c2 \cdot s} = \bar{y}_r\left[1 + \left(1 - \frac{m}{n}\right)\left\{\frac{\bar{x}_o^2}{\bar{x}_r\bar{x}_s} - 1\right\}\right]. \quad (2.10)$$

The correction factors given in (2.8) are not ideal when the correlation between x and y is close to 1. In this case, we have $\hat{B}_r \approx \hat{B}_s \approx \hat{B}_o$, provided that the model (2.2) holds. Therefore, the correction factor C should be close to 1. However, the correction factors given in (2.8) could be very different from 1 and using them would bring bias. For this reason, it may be preferable to work with a correction factor C in (2.7) that takes the correlation into account. Correction factors of this kind are

$$k_i = 1 - \{(c_i^2 - 1)(\hat{R}_{xy}^2 - 1)\}, \quad (2.11)$$

where c_i, $i = 1, \ldots, 4$, are the four correction factors given in (2.8), and $\hat{R}_{xy}$ is the estimated correlation coefficient based on the respondent data. In our Monte Carlo simulation we also included the estimator (2.7) corresponding to the four choices $C = k_i$, $i = 1, \ldots, 4$. These estimators will be denoted as $\bar{y}_{k_i \cdot s}$, $i = 1, \ldots, 4$.

3. VARIANCE ESTIMATION

Since we are interested in variance estimators based on single value imputation, the variance estimation method proposed in Särndal (1990, 1992) is of interest. Assuming unconfounded nonresponse and that the model ξ in (2.3) holds, the variance estimator for the point estimator $\bar{y}_{raimp}$ in (2.4) obtained by this method is given by

$$\hat{V}(\bar{y}_{raimp}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2}{n - 1} + \left(\frac{1}{n} - \frac{1}{N}\right)A_o\hat{\sigma}^2 + \left(\frac{1}{m} - \frac{1}{n}\right)A_1\hat{\sigma}^2$$

$$= \hat{V}_{ord} + \hat{V}_{dif} + \hat{V}_{imp}, \quad (3.1)$$

where

$$A_o = \frac{1}{n - 1}\left\{\sum_o x_k - \frac{\sum_o x_k^2}{\sum_r x_k} + \frac{\bar{x}_s \sum_o x_k}{\sum_r x_k}\right\},$$

$$A_1 = \frac{\bar{x}_s\bar{x}_o}{\bar{x}_r}$$

and

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2/(m - 1)}{\bar{x}_r[1 - (cv_{xr})^2/m]}, \quad (3.2)$$

where

$$e_k = y_k - \hat{B}_r x_k, \; cv_{xr} = \frac{\sqrt{\sum_r (x_k - \bar{x}_r)^2/(m - 1)}}{\bar{x}_r}.$$

The variance of $\bar{y}_{raimp}$ has two components, namely, the sampling variance and the variance due to imputation. The first term in (3.1) (denoted by $\hat{V}_{ord}$) is an estimate of the sampling variance calculated using the ordinary variance formula assuming that imputed data are as good as real observations. Since this assumption does not hold, $\hat{V}_{ord}$ underestimates the true sampling variance. To correct this underestimation, the second term $\hat{V}_{dif}$ in (3.1) is added. The last term $\hat{V}_{imp}$ in (3.1) is an estimate of the variance due to imputation.

If we compute the mean of the y-values from the completed data set $\{y_k^c : k \in s\}$ given in (2.6), we get the estimator (2.7). Its variance estimator should take the correction factor C into account. If we can assume that the expectation $E_\xi E_p E_q$ is equal to $E_p E_q E_\xi$ (this is true under unconfounded nonresponse), we can use Särndal's (1990, 1992) method to obtain a variance estimator which takes C into account. However, we are mainly interested in confounded cases. We are therefore proposing a variance estimator based on the following heuristic argument.

The estimator $\hat{\sigma}^2$ in (3.2) uses the respondent data only. It will certainly be biased for confounded mechanisms and some correction is needed in order to use formula (3.1) for the corrected estimator (2.7). We suggest to replace $\hat{\sigma}^2$ in (3.1) by $C^2\hat{\sigma}^2$, to obtain the following variance estimator for the estimator $\bar{y}_{c\cdot s}$ in (2.7):

$$\hat{V}(\bar{y}_{c\cdot s}) = \hat{V}^c_{\text{ord}} + C^2(\hat{V}_{\text{dif}} + \hat{V}_{\text{imp}}), \qquad (3.3)$$

where $\hat{V}^c_{\text{ord}}$ is computed using the data after imputation with the bias correction factor C. Replacing C^2 by c_i^2 or k_i^2, we obtain the variance estimators corresponding to $\bar{y}_{ci\cdot s}$ or $\bar{y}_{ki\cdot s}$. The resulting variance estimators work quite well in many of the cases covered in the simulation reported in Section 5.

4. SIMULATION STUDY

We are considering eight corrected estimators corresponding to the eight correction factors given in (2.8) and (2.11). A simulation study was conducted to determine whether the corrected estimators succeed in restoring $\bar{y}_s$ under different response mechanisms, in particular, confounded mechanisms. For comparison, we also included the uncorrected estimators $\bar{y}_r$ and $\bar{y}_{\text{raimp}} = \bar{x}_s\bar{y}_r/\bar{x}_r$ given by (2.2). Our primary objective was to examine the corrected estimators when the finite population follows the ratio model ξ given by (2.3). However, we also wanted to see how the corrected estimators behave under relationships other than linear regression through the origin.

We also studied the coverage rates associated with the different estimators when the confidence intervals are computed with the aid of the variance estimators proposed in Section 3.

For the simulation, we generated 12 different finite populations, each of size $N = 100$, by specifying in different ways the constants a, b, c, and d in the regression model:

$$\Xi: y_k = a + bx_k + cx_k^2 + \epsilon_k, \quad E_\Xi(\epsilon_k) = 0,$$
$$V_\Xi(\epsilon_k) = d^2x_k, \qquad (4.1)$$

where the ϵ_k are assumed to be independent. Four different regression types were created by four different specifications of (a, b, c). These types are called RATIO ($a = c = 0, b > 0$, thus conforming to the ratio model ξ in (2.3)), CONCAVE ($a = 0, b > 0, c < 0$), CONVEX ($a = 0, b > 0, c > 0$) and NONRATIO ($a \neq 0, b > 0, c = 0$). For each regression type, three different levels of the model correlation ρ_{xy}, 0.7, 0.8 and 0.9, were obtained by a suitable choice of d. This resulted in 12 specifications of (a, b, c, d) as shown in Table 1.

Table 1

Characteristics of the Populations

POP	TYPE	a	b	c	d	R_{xy}	MEAN of y
1	RATIO	0	1.5	0	6.12	0.69	70.95
2	RATIO	0	1.5	0	4.50	0.81	69.92
3	RATIO	0	1.5	0	2.91	0.90	72.67
4	CONCAVE	0	3	−0.01	6.78	0.71	117.27
5	CONCAVE	0	3	−0.01	4.83	0.81	114.57
6	CONCAVE	0	3	−0.01	2.80	0.90	112.11
7	CONVEX	0	0.25	0.01	5.98	0.71	35.89
8	CONVEX	0	0.25	0.01	4.22	0.81	37.06
9	CONVEX	0	0.25	0.01	2.35	0.90	43.92
10	NON-RATIO	20	1.5	0	6.12	0.71	95.25
11	NON-RATIO	20	1.5	0	4.50	0.81	94.46
12	NON-RATIO	20	1.5	0	2.91	0.90	93.32

For each of the 12 specifications, we generated 100 population values (y_k, x_k), $k = 1, \ldots, 100$, by a two step process. We used the Γ-distribution with parameters α and β. Its density is

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}\exp(-x/\beta) \quad \text{for} \quad x > 0. \qquad (4.2)$$

First, we generated 100 values x_k, $k = 1, \ldots, 100$, according to the Γ-distribution with parameters $\alpha = 3$, $\beta = 16$, implying that the mean is $\alpha\beta = 48$ and the variance $\alpha\beta^2 = 768$. Then, for each fixed x_k, $k = 1, \ldots, 100$, we generated one value y_k according to the Γ-distribution with parameters

$$\alpha = \frac{\{\mu(x)\}^2}{\sigma^2(x)} = \frac{(a + bx + cx^2)^2}{d^2x}, \qquad (4.3)$$

$$\beta = \frac{\sigma^2(x)}{\mu(x)} = \frac{d^2x}{a + bx + cx^2}, \qquad (4.4)$$

where $x = x_k$ and (a, b, c, d) is one of the 12 vectors fixed in advance. This implies that $E_\Xi(y_k \mid x_k) = \alpha\beta = a + bx_k + cx_k^2$ and $V_\Xi(y_k \mid x_k) = \alpha\beta^2 = d^2x_k$, as required under the model (4.1). The same x-values were used for all 12 populations. For the populations generated by this process, Table 1 shows the values of the population correlation R_{xy} and the population mean of y. Note that the values of a, b, c, and d were chosen so as to obtain realistic types of populations that can be encountered in practice.

To simulate nonresponse, we used five different nonresponse mechanisms, each defined by independent Bernoulli (Θ_k) trials, where the probability of nonresponse Θ_k for unit k was specified as follows:

(M1) Θ_k is constant and independent for all $k \in U$. This is the uniform response mechanism, therefore unconfounded.

(M2) Θ_k is a decreasing function of x_k specified as $\Theta_k = \exp(-\gamma x_k)$. This is an unconfounded mechanism.

(M3) Θ_k is an increasing function of x_k specified as $\Theta_k = 1 - \exp(-\gamma x_k)$. This is also an unconfounded mechanism.

(M4) Θ_k is a decreasing function of y_k specified as $\Theta_k = \exp(-\gamma y_k)$. This is a confounded mechanism.

(M5) Θ_k is an increasing function of y_k specified as $\Theta_k = 1 - \exp(-\gamma y_k)$. This is also a confounded mechanism.

Note that since we assume x and y to be positively correlated, both (M2) and (M4) are mechanisms such that large units respond more often than small units. The smaller units will be underrepresented in the response set r. Conversely, (M3) and (M5) are mechanisms such that small units respond more often than large units. The larger units will be underrepresented in the response set r.

The first mechanism corresponds to the naive Assumption I discussed in Section 2. (M2) and (M3) correspond to Assumption II while (M4) and (M5) represent fairly simple examples of the confounded mechanisms discussed in connection with Assumption III. For (M2), (M3), (M4) and (M5), the constant γ was determined in such a way that the average nonresponse probability $\bar{\Theta} = (1/N) \sum_U \Theta_k$, is equal to one of the values 10%, 20%, 30% and 40%. Therefore, for each population, there were $5 \times 4 = 20$ different combinations of nonresponse mechanism and nonresponse rate.

For each of the 12 populations, 1,000 samples of size $n = 30$ were drawn. Then for each realized sample, 50 response sets were generated using independent Bernoulli(Θ_k) trials according to one of the 20 combinations of nonresponse mechanism and nonresponse rate. Thus 50,000 response sets were realized for each of the $12 \times 20 = 240$ combinations resulting from cross-classifying the 12 populations with the 20 combinations of nonresponse mechanism and nonresponse rate.

5. RESULTS

We studied the two uncorrected estimators $\bar{y}_r$ (justified under Assumption I) and $\bar{y}_{\text{raimp}} = \bar{x}_s \bar{y}_r / \bar{x}_r$ (justified under Assumption II) and the 8 corrected estimators $\bar{y}_{ci \cdot s}$ and $\bar{y}_{ki \cdot s}$, $i = 1, \ldots, 4$ (justified under Assumption III). (We call both $\bar{y}_r$ and $\bar{y}_{\text{raimp}}$ uncorrected even though (2.5) shows that we can view $\bar{y}_{\text{raimp}}$ as a corrected version of the naive estimator $\bar{y}_r$. Recall that our principal aim is to correct the bias of $\bar{y}_{\text{raimp}}$ when the mechanism is confounded.)

The performance of the 10 estimators is judged by the magnitudes of the relative bias (RB), the relative root mean square error (RRMSE), and the coverage rate (CVR). The RB and the RRMSE of a point estimator $\hat{\bar{y}}_U$ for $\bar{y}_U$ are defined respectively as,

$$\text{RB}(\bar{y}) = 100 \times \frac{E_p E_q(\hat{\bar{y}}_U) - \bar{y}_U}{\bar{y}_U},$$

$$\text{RRMSE}(\bar{y}) = 100 \times \frac{\sqrt{E_p E_q(\hat{\bar{y}}_U - \bar{y}_U)^2}}{\bar{y}_U}.$$

The expectations $E_p E_q(\hat{\bar{y}}_U)$ and $E_p E_q(\hat{\bar{y}}_U - \bar{y}_U)^2$ were estimated by Monte Carlo simulation using the 50,000 realized response sets for each of 240 combinations. With this number of replicates, the Monte-Carlo error was less than 0.1%, assuming that the distribution of the $\hat{\bar{y}}_U$'s is approximately normal. We will use the abbreviation ARB to denote the absolute relative bias, $|\text{RB}(\bar{y})|$.

We will also discuss the coverage rate (CVR) of the 95% confidence interval constructed as

$$\hat{\bar{y}}_U \pm 1.96\sqrt{\hat{V}(\hat{\bar{y}}_U)}, \tag{5.1}$$

where $\hat{\bar{y}}_U$ is one of the 10 estimators and $\hat{V}(\hat{\bar{y}}_U)$ the corresponding variance estimator. For $\bar{y}_{\text{raimp}}$ and the 8 corrected estimators, we used the variance estimators described in Section 3. For $\bar{y}_r$, we used the variance estimator

$$\hat{V}(\bar{y}_r) = \left(\frac{1}{m} - \frac{1}{N}\right) \sum_r (y_k - \bar{y}_r)^2/(m-1).$$

The CVR is calculated as 100 times the proportion of the 50,000 response sets such that the interval computed in the manner of (5.1) includes the true mean $\bar{y}_U$.

For the following discussion, we group the corrected estimators into two groups: s-corrected estimators, which are based on correction factors involving $\bar{x}_s$ or $\bar{w}_s$, that is, c_2, c_4, k_2 and k_4 and r-corrected estimators, which are based on correction factors involving $\bar{x}_r$ or $\bar{w}_r$, that is, c_1, c_3, k_1 and k_3.

The nonresponse mechanism is the key to the performance of the various estimators. Therefore, Tables 2 and 3 show the behavior of the estimators separately for each of the five mechanisms. We noted that the correlation level and the nonresponse rate do not have a very pronounced effect on the ranking of the estimators. Thus the performance measures ARB, RRMSE and CVR were averaged over 12 cases (three correlation levels $\times$ four nonresponse rates). These averages are shown in Table 2 for the RATIO type regression and in Table 3 for the CONCAVE, CONVEX and NONRATIO regression types.

Table 2

Average ARB, RRMSE (RM) and CVR of Ten Different Estimators for the RATIO Type Populations

For each mechanism, 12 cases were averaged (four nonresponse rates × three correlation levels)

	M1 (uniform)			M2 (decreasing-x)			M3 (increasing-x)			M4 (decreasing-y)			M5 (increasing-y)		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
$\bar{y}_r$	0.2	13.9	92.5	12.9	19.1	86.0	9.5	16.5	81.1	19.1	23.6	72.3	14.9	19.9	68.2
$\bar{y}_{\text{raimp}}$	0.2	12.3	92.7	0.6	11.8	93.0	0.4	12.9	92.4	5.3	13.0	92.5	6.0	13.9	85.6
$\bar{y}_{c2\cdot s}$	1.0	13.3	92.4	4.4	12.6	88.9	8.9	18.3	93.0	1.8	11.8	92.4	3.6	15.3	92.2
$\bar{y}_{c4\cdot s}$	0.9	13.2	92.3	4.7	12.6	88.6	8.4	17.7	93.0	1.7	11.7	92.3	3.4	14.9	92.2
$\bar{y}_{k2\cdot s}$	1.1	13.2	92.8	2.4	12.0	90.9	8.0	18.5	93.5	1.7	11.7	93.3	2.2	15.3	92.0
$\bar{y}_{k4\cdot s}$	1.0	13.1	92.7	2.6	12.0	90.8	7.3	17.7	93.5	1.6	11.7	93.2	1.8	14.7	91.9
$\bar{y}_{c1\cdot s}$	1.7	14.7	91.4	5.9	13.4	86.4	15.7	26.2	87.6	1.9	12.2	90.9	8.9	21.3	89.8
$\bar{y}_{c3\cdot s}$	1.6	14.4	91.4	6.2	13.5	86.1	14.9	25.1	87.8	2.1	12.2	90.7	8.3	20.4	90.0
$\bar{y}_{k1\cdot s}$	2.0	14.7	92.3	3.1	12.3	90.0	15.9	29.6	88.9	1.1	11.7	92.8	8.3	23.8	90.7
$\bar{y}_{k3\cdot s}$	1.7	14.3	92.3	3.2	12.4	89.8	14.6	27.6	89.3	1.0	11.7	92.7	7.1	21.9	91.0

Table 3

Average ARB, RRMSE (RM) and CVR of Six Different Estimators for CONCAVE, CONVEX, and NONRATIO Populations

(For each mechanism, 12 cases are averaged as in Table 2)

	M1			M2			M3			M4			M5		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
							CONCAVE								
$\bar{y}_r$	0.2	10.4	92.9	10.5	14.8	82.3	7.3	12.7	82.3	12.3	16.0	78.3	8.7	13.4	78.8
$\bar{y}_{\text{raimp}}$	0.2	9.4	94.5	1.4	9.1	93.4	2.6	10.5	94.9	1.9	9.2	94.9	2.1	9.7	92.9
$\bar{y}_{c2\cdot s}$	1.1	11.4	92.4	6.3	11.4	84.7	11.8	18.8	88.4	3.2	10.2	90.0	5.5	14.2	92.3
$\bar{y}_{c4\cdot s}$	1.0	11.1	92.8	6.6	11.5	84.3	11.4	18.0	88.8	3.6	10.3	89.8	5.5	13.7	92.7
$\bar{y}_{k2\cdot s}$	1.0	10.7	93.7	4.5	10.1	89.1	9.5	16.8	91.6	1.7	9.3	93.0	3.7	12.8	93.7
$\bar{y}_{k4\cdot s}$	0.9	10.5	93.8	4.6	10.1	89.0	9.0	16.0	91.8	1.8	9.3	92.8	3.5	12.3	93.9
							CONVEX								
$\bar{y}_r$	0.9	23.7	90.9	19.0	31.6	92.3	15.0	26.5	76.1	33.2	41.7	76.4	37.1	41.4	37.5
$\bar{y}_{\text{raimp}}$	0.6	21.4	90.6	5.8	21.7	92.8	7.0	22.1	85.6	14.0	25.0	90.0	27.6	33.5	52.0
$\bar{y}_{c2\cdot s}$	1.2	21.1	91.8	0.4	19.8	91.8	2.0	22.2	92.4	7.3	20.8	93.4	17.8	28.2	71.7
$\bar{y}_{c4\cdot s}$	1.2	21.3	91.5	0.3	19.9	91.5	1.8	22.3	92.4	6.7	20.6	93.4	18.5	28.5	70.5
$\bar{y}_{k2\cdot s}$	1.6	21.2	91.9	3.0	21.0	92.0	3.0	22.2	92.6	9.8	22.7	91.7	16.2	27.6	74.0
$\bar{y}_{k4\cdot s}$	1.4	21.3	91.6	2.9	21.0	91.8	2.6	22.0	92.3	9.5	22.7	91.7	17.6	27.7	72.6
							NON-RATIO								
$\bar{y}_r$	0.1	10.7	92.9	9.7	14.6	86.5	7.3	12.6	81.3	11.9	16.1	80.8	8.8	13.5	77.8
$\bar{y}_{\text{raimp}}$	0.2	9.6	94.5	2.1	9.5	92.4	2.6	10.5	95.3	2.1	9.6	94.4	1.6	9.9	93.3
$\bar{y}_{c2\cdot s}$	1.1	11.4	92.5	7.0	11.9	83.5	11.9	18.8	89.2	2.6	10.0	90.9	5.3	14.5	92.5
$\bar{y}_{c4\cdot s}$	1.0	11.3	92.4	7.3	12.1	82.8	11.5	18.1	89.4	2.7	10.1	90.6	4.9	13.8	92.7
$\bar{y}_{k2\cdot s}$	1.3	11.2	93.4	5.0	10.9	86.9	11.3	19.0	90.7	1.3	9.6	92.8	4.7	14.3	93.5
$\bar{y}_{k4\cdot s}$	1.1	10.9	93.4	5.2	11.1	86.5	10.6	17.8	91.1	1.3	9.7	92.6	4.1	13.4	93.8

We now comment on the tables. A conclusion of general character is that the respondent mean $\bar{y}_r$ has, as expected, a large bias and a very poor CVR for all of the nonuniform mechanisms. Its performance is satisfactory only for the uniform mechanism (M1). Thus we can focus on the comparisons between the uncorrected $\bar{y}_{\text{raimp}}$ on the one hand and the eight corrected estimators on the other. For both of the criteria ARB and RRMSE, we noted that the *s*-corrected estimators generally gave better results than the *r*-corrected ones. This is clearly seen in Table 2, where *s*-corrected and *r*-corrected estimators are displayed in two separate groups. Given this better behavior of the *s*-corrected group, we deleted the *r*-corrected group in Table 3.

5.1 RATIO Type Regression

From Table 2, we draw the following conclusions.

(i) The mechanism (M1) (uniform nonresponse).

When the mechanism (M1) holds, the uncorrected estimator $\bar{y}_{\text{raimp}}$ is essentially bias free, and there is no need to correct. However, if the analyst, suspecting a confounded mechanism, has nevertheless chosen one of the corrected estimators, the penalty is not severe. The eight corrected estimators show only a small increase in ARB and in RRMSE compared to $\bar{y}_{\text{raimp}}$.

(ii) The mechanisms (M2) and (M3) (unconfounded, nonuniform and *x*-value dependent).

For these mechanisms, the ARB is seen to be very small for the uncorrected estimator $\bar{y}_{\text{raimp}}$, as theory would lead us to expect. Our interest is instead focused on the behavior of the eight corrected estimators, since it is important to know if a penalty is associated with an incorrect decision to use one of these estimators. Such a decision would be brought about by an incorrect assumption that the response mechanism is confounded (when in fact it is unconfounded but nonuniform). Table 2 shows that there is indeed some penalty in the form of both increased ARB and increased RRMSE. The penalty is less severe for the *s*-corrected group. For both groups, the penalty is less severe for the mechanism (M2) than for the mechanism (M3).

(iii) The mechanism (M4) (confounded and *y*-value dependent).

For this mechanism, a striking feature of Table 2 is that all eight corrected estimators give a substantial bias reduction compared to the uncorrected estimator $\bar{y}_{\text{raimp}}$ (and a very large reduction relative to the naive estimator $\bar{y}_r$). The corrected estimators also show some improvement in RRMSE compared to $\bar{y}_{\text{raimp}}$. The *s*-corrected estimators perform better than the *r*-corrected ones. Within the *s*-corrected group of estimators, the differences are minor, as is the case within the *r*-corrected group.

(iv) The mechanism (M5) (confounded and *y*-value dependent).

Table 2 shows that the *s*-corrected estimators have a smaller ARB than the uncorrected $\bar{y}_{\text{raimp}}$; their RRMSE is slightly higher. By contrast, the *r*-corrected estimators "overcorrect" so that both the ARB and the RRMSE exceed the levels observed for $\bar{y}_{\text{raimp}}$. The *r*-corrected group does not perform well for this mechanism.

In summary, Table 2 shows that if the ratio model (2.2) holds and the assumption of a confounded mechanism is correctly made, the decision to use one of the corrected estimators may lead to a reduced bias. The main difficulty facing the analyst is to accurately predict the nature of the response mechanism causing nonresponse. In particular, it may be difficult for the analyst to separate a confounded mechanism (*e.g.*, one with $\Theta_k = e^{-\gamma y_k}$) from a similar nonuniform unconfounded mechanism (*e.g.*, one with $\Theta_k = e^{-\gamma x_k}$). Yet this subtle difference has a marked effect on the bias of $\bar{y}_{\text{raimp}}$ and on the decision whether or not to use a corrected estimator. When the nonuniform unconfounded type applies, we have seen that there is a penalty associated with the corrected estimators, in particular with the *r*-corrected group.

5.2 Other Regression Types

Table 3 shows the performance of six estimators (the two uncorrected and the four *s*-corrected) for the CONCAVE, CONVEX, and NONRATIO regression types. As in Table 2, there is little to choose between the estimators when the uniform mechanism (M1) holds. For the two confounded mechanisms, the results in Table 3 do not send a clear message that *s*-corrected estimation should be attempted even if the assumption of a confounded mechanism is correctly made. Compared to the uncorrected $\bar{y}_{\text{raimp}}$, the *s*-corrected estimators show a clearly improved performance (in terms of smaller ARB and smaller RRMSE) only for the CONVEX population type. Even in this case, a substantial bias remains after the attempt at correction. For the two unconfounded nonuniform mechanisms (M2) and (M3), it is *a priori* clear that one would not expect improved performance on the part of the *s*-corrected estimators when compared to $\bar{y}_{\text{raimp}}$. Oddly enough however, we find that the *s*-corrected estimators work very well for the CONVEX population. These conclusions leave the analyst with a difficult choice if a RATIO type population cannot be assumed. Then it is difficult on the basis of our findings to recommend the use of one of the corrected estimators.

5.3 Coverage Rates

Tables 2 and 3 also show that the variance estimation procedure suggested in Section 3 generally works well. Indeed the coverage rates for the corrected estimators are uniformly good whenever the ARB is small. In particular,

Table 4

Average ARB, RRMSE (RM) and CVR of the Two Uncorrected Estimators and the c_4 - and k_4 - Corrected Estimators
(Averaged Over All Population Types)

	M1			M2			M3			M4			M5			Overall		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
$\bar{y}_r$	0.3	14.7	92.3	13.0	20.0	86.8	9.8	17.1	80.2	19.1	24.4	77.0	17.4	22.1	65.6	11.9	19.6	80.4
$\bar{y}_{\text{raimp}}$	0.3	13.2	93.1	2.5	13.0	92.9	3.1	14.0	92.0	5.8	14.2	93.0	9.3	16.7	81.0	4.2	14.2	90.4
$\bar{y}_{c4\cdot s}$	1.0	14.2	92.3	4.7	14.0	86.8	8.3	19.0	90.8	3.7	13.2	91.5	8.1	17.7	87.0	5.2	15.6	89.7
$\bar{y}_{k4\cdot s}$	1.1	14.0	92.9	3.8	13.6	89.5	7.4	18.4	92.2	3.6	13.3	92.6	6.7	17.0	88.0	4.5	15.2	91.0

for the unconfounded mechanisms (M2) and (M3), the coverage rates for the corrected estimators are about equal to or better than those for the uncorrected estimators.

5.4 Overall Comments

From the summary Table 4, we note that, as expected, $\bar{y}_r$ and $\bar{y}_{\text{raimp}}$ show the best performance for the uniform response mechanism (M1). The uncorrected estimator $\bar{y}_{\text{raimp}}$ is the best one for the unconfounded mechanisms (M2) and (M3), while the corrected estimators are the best ones for the confounded mechanism (M4) and (M5).

Finally, on the average over all 240 cases included in our study, we note from the overall column of Table 4 that $\bar{y}_{\text{raimp}}$ and $\bar{y}_{k4\cdot s}$ perform similarly with the former having a slightly smaller bias and the latter having slightly better coverage rate.

6. CONCLUSIONS

It has long been recognized that nonresponse causes bias in survey estimates, except in rare cases. Imputation is a widely used practice to handle nonresponse, because it is convenient to work with a complete data set. There are many imputation rules as well as some softwares that can be used in large scale surveys. Imputation is sometimes applied without critical questioning, and, although widely used, imputation does not solve the critical problem of bias caused by nonresponse.

In this paper, we have examined ratio imputation. The ordinary ratio imputation $\hat{B}_r x_k$ is justified (that is, it produces no bias) if two conditions hold: (a) the regression model behind the ratio imputation rule holds (that is, a linear regression through the origin); (b) the response mechanism is unconfounded.

The results of our simulation give some idea of the magnitude of the bias of the usual ratio imputation estimator $\bar{y}_{\text{raimp}}$ when one or both of the two conditions break down. We considered several nonuniform response mechanisms, confounded as well as unconfounded mechanisms. We also considered breakdown of the regression model behind ratio imputation.

We argued that a confounded mechanism can sometimes be realistically assumed in a survey. We showed that if an assumption of confounded response mechanism is correctly made, and if the model behind the ratio imputation is valid, one can make some progress toward bias reduction using the s-corrected estimators in this paper. They have substantially less bias than the uncorrected estimator $\bar{y}_{\text{raimp}}$. The s-corrected estimators are generally more effective than the r-corrected estimators for reducing the bias.

Suppose the analyst is working under the assumption that the ratio model (2.2) holds. Our simulation study then leads to suggested estimators according to the following Table 5, depending on the assumed nature of the response mechanism and on the nonresponse rate. The entry "any" means any of the 10 estimators in Table 2.

Table 5

Suggested Estimators for Each Nonresponse Mechanism

Nonresponse Rate	Suggested Estimator: Response Mechanism		
	Uniform	Unconfounded	Confounded
($\leq$ 10%)	any	any but $\bar{y}_r$	any but $\bar{y}_r$
($>$ 10%)	any[1]	$\bar{y}_{\text{raimp}}$	s-corrected

Note 1: $\bar{y}_{\text{raimp}}$ as a slight advantage over the others.

If the regression model behind ratio imputation fails, the situation is less clear. Unless the naive assumption of a uniform response mechanism holds (which is unlikely), the uncorrected ratio imputation estimator $\bar{y}_{\text{raimp}}$ can have considerable bias. We found that $\bar{y}_{\text{raimp}}$ is particularly prone to bias for the CONVEX type population where the s-corrected group of estimators usually have smaller bias than $\bar{y}_{\text{raimp}}$. On the other hand, for the CONCAVE and the NONRATIO type populations, $\bar{y}_{\text{raimp}}$ is generally more resistant to bias than the s-corrected estimators.

7. ACKNOWLEDGMENT

The authors wish to thank the referees and the associate editor for their helpful comments. An earlier version of this paper was presented at the Annual Research Conference (ARC) in Arlington, Virginia, March 22-25, 1992.

REFERENCES

BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

FAY, R.E. (1989). Estimating nonignorable nonresponse in longitudinal surveys through causal modeling. In *Panel Surveys* (Eds. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh), 375-399.

GREENLESS, J.S., REECE, W.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 82, 251-261.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

RUBIN, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.

RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, 337-347.

SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 149–158
Statistics Canada

Dual System Estimation of Census Undercount in the Presence of Matching Error

YE DING and STEPHEN E. FIENBERG[1]

ABSTRACT

Dual system estimation (DSE) has been used since 1950 by the U.S. Bureau of Census for coverage evaluation of the decennial census. In the DSE approach, data from a sample is combined with data from the census to estimate census undercount and overcount. DSE relies upon the assumption that individuals in both the census and the sample can be matched perfectly. The unavoidable mismatches and erroneous nonmatches reduce the accuracy of the DSE. This paper reconsiders the DSE approach by relaxing the perfect matching assumption and proposes models to describe two types of matching errors, false matches of nonmatching cases and false nonmatches of matching cases. Methods for estimating population total and census undercount are presented and illustrated using data from 1986 Los Angeles test census and 1990 Decennial Census.

KEY WORDS: Capture-recapture; Matching bias; Modelling matching error; Multinomial likelihood.

1. INTRODUCTION

The problem of undercount in the U.S. census has been of special concern since the first census of 1790 (Jefferson 1986). The DSE (or capture-recapture) approach has been used in conjunction with the census to evaluate population coverage as part of what is called the post-enumeration survey (PES) program. Ericksen and Kadane (1985) and Wolter (1986) describe the use of the DSE approach in the context of the 1980 decennial census. A new design for the PES was planned for the 1990 decennial census and refinements in methodology were examined in connection with a 1986 test census in central Los Angeles County, referred to as the Test of Adjustment Related Operations (TARO). Diffendal (1988) discusses methodology, operations, and the results of TARO, and Hogan and Wolter (1988) and Schenker (1988) provide evaluation of the operations and assumptions underlying the DSE approach.

The PES approach to dual-system estimation uses two samples, called the P-sample and the E-sample. The P-sample which is drawn separately from the census, helps to measure census omissions; the E-sample drawn from the census enumerations, helps to measure census erroneous enumerations. For the 1986 TARO, the dual-system estimator for the population size, N, which combines the information from the P-sample and the E-sample takes the form:

$$\hat{N} = (\text{CEN} - \text{EE} - \text{SUB}) \cdot N_p/M,$$

where CEN is the unadjusted census count; EE is the estimated number of erroneous enumerations and unmatchable persons included in the census; SUB is the number of whole-person substitutions in the census; N_p is the number of people in the P-sample; M is the estimate of the number of people in both census and the P-sample. For details see Diffendal (1988) or Wolter (1986). For the variation on this formula as used in conjunction with the 1990 census, see Hogan (1992, 1993).

DSE and the matching problem gained considerable attention in the 1970's due to its use in estimating births and deaths in developing countries, and it is thought by some that perhaps the greatest problem with the dual-system estimation approach used in 1980 census was the rate of matching error (Fienberg 1989). Jaro (1989) describes the technological innovations for matching introduced by the Bureau of the Census for 1990 and the test of the related matching methodology in a 1985 pretest. Biemer (1988) considers models for evaluating the impact of matching error on estimates of census coverage error without attempting to correct for the matching bias in the usual dual-system estimate. The actual procedure used in the 1990 census included not only a computer matching algorithm and various clerical follow-ups but also logistic regression models for unresolved cases in both the P-sample and E-sample (see Belin *et al.* 1993).

Matching is used to determine the census enumeration status of the people enumerated in the P-sample. Specifically, those people in the P-sample who are matched to the census are considered to have been enumerated. People in the P-sample who do not match are, for the most part, considered to have been missed by the census. Matching errors can occur for two general reasons:

[1] Ye Ding is Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

1. The information reported by the respondents/interviewers was incorrect.
2. Correct information was reported, but it was not correctly used.

Moreover, two types of errors can occur: false matches of nonmatching cases and false nonmatches of matching cases. False matches of nonmatching cases may be divided into

(a) instances in which a P-sample case was erroneously matched to the enumeration of another person, but a match to that actual E-sample case should have been made, and
(b) instances in which no match should have been made.

The former case is not "serious" for the purposes of estimating N, since such false matches would have been, in fact, correctly classified as a match to the census. In the second case, however, the number of nonmatches becomes understated. False nonmatches to the census, on the other hand, have the effect of overestimating the nonmatch rate. Fay, Passel, Robinson and Cowan (1988) note that false nonmatches probably represent a greater concern than false matches. False matches are less common than false nonmatches because matches can be reviewed easily.

In Section 2, we propose models for matching errors and then, in Section 3 and 4, we present a systematic procedure for the estimation of the population total and thus the census undercount. In Section 5, we analyze the data from 1986 Los Angeles test census and 1990 Decennial Census to show how our method accounts for matching errors in the undercount estimates.

2. MODELING MATCHING ERRORS

For simplicity, we assume that the matching mechanism is constrained, in the sense that no individual in one sample can be matched with more than one individual in another sample. Moreover, we implicitly assume a version of simple random sampling, within strata, and this yields a standard multinomial sampling model for dual system estimation. This simplification allows us to focus on the impact of matching and its mechanisms. In what follows, we provide a way to view the recapture data, for the purpose of setting up models for matching.

Let $Z_{N\times 1}$ be the characteristic vector for the whole population, such that the i-th component of $Z_{N\times 1}$ contains the characteristics for the i-th individual, where $1 \le i \le N$. Not all the components in $Z_{N\times 1}$ can be observed in any one sample. The object is to estimate N, the size of the population, from information from two samples. One could view drawing a sample from the population as drawing some components in $Z_{N\times 1}$ at random to form a new vector Y. Then, missing or misreporting of certain characteristics in those components drawn may cause matching errors. Henceforth we will refer to the first sample as Y_1 and the second sample as Y_2, and in the following discussion they will be the two capture-recapture samples for dual system estimation.

Two types of matching errors can occur: false nonmatches of matching cases, and false matches of nonmatching cases. We will refer to the former as a type 1 error and the latter as a type 2 error. We can focus on modeling one or both types of error. Under perfect matching, each component in Y_1 or Y_2 contains the same information as in $Z_{N\times 1}$, and the number of matches will be the number of elements common to Y_1 and Y_2. When faced with uncertain matching, we consider the following simple model:

Model (A):

(i) Assume that those matched pairs of components under perfect matching will still be matched, each with common probability α, $0 < \alpha \le 1$.
(ii) All those unmatched will remain unmatched, *i.e.*, no false matches.

Model (A) characterizes a mechanism for type 1 matching error with error probability $1 - \alpha$, assuming that type 2 matching error is negligible.

To develop a model for both types of matching error, we need to consider carefully all the possibilities that lead to false matches. When there is no matching error, one can write $Y_1 = (M_1, N_1)$ and $Y_2 = (M_2, N_2)$, so that sets M_1 and M_2 have the same size and every individual in M_1 is correctly matched with one individual in M_2 and vice versa, N_1 is the set of those in sample Y_1 who are not matched with any one in sample Y_2, and N_2 is the set of those in sample Y_2 who are not matched with any one in sample Y_1. When matching errors are present, false matches can occur in the following ways:

(a) A person in M_1 is matched incorrectly with a person in M_2.
(b) A false match occurs between M_1 and N_2.
(c) A false match occurs between M_2 and N_1.
(d) A false match occurs between N_1 and N_2.

We note that each of (a), (b), (c) happens only when at least 2 errors are made, that is, the correct match is not made and an incorrect match is made. Since such errors occur with small probability, we assume for simplicity that cases (a), (b), (c) have negligible probability of occurrence in the next model.

Model (B):

(i) Assume, as in model (A), that matching pairs between M_1 and M_2 will still be matched, but with probability α, $0 < \alpha \le 1$.
(ii) Assume that false matches of types (a), (b), (c) are negligible.
(iii) Assume that each person in N_1 will be matched with someone in N_2 with a common probability β, $0 \le \beta < 1$.

Even though, in theory, both α and β can vary from 0 to 1, in the census context we expect that $\alpha \approx 1$, and $\beta \approx 0$.

We can also consider instances in which the matching error probabilities and capture probabilities potentially vary over identifiable population subgroups. In other words, the population can be divided into strata, by demographic (*e.g.*, age, race, sex) and geographic variables, within which the matching error probabilities and capture probabilities could be assumed to be more homogeneous than in the whole population. Suppose the whole population consists of I strata. Let $Z^i_{N_i \times 1}$ be the characteristic vector for the population of the i-th stratum with unknown size N_i, and let Y_{i1}, Y_{i2} be two samples taken from the i-th stratum which are used to get an estimate $\hat{N}_i$. Then we can form an estimate of the overall population size by setting $\hat{N} = \sum_{i=1}^{I} \hat{N}_i$. We can refine models (A) and (B) as follows:

Model (A′):

Assume model (A) holds within each stratum, and let α_i be the probability of a match for matching components in stratum i, $0 < \alpha_i \leq 1$, $1 \leq i \leq I$.

Model (B′):

Assume model (B) holds within each stratum, and let the two probability parameters for i-th stratum be α_i, β_i, $1 \leq i \leq I$.

For 1990 PES, the P-sample matching was conducted using the sample blocks plus a ring of surrounding blocks (Hogan 1993). Geocoding errors may lead to false matches across geographically defined post-strata, and false matches are possible for demographically defined post-strata. Models (B′) implicitly assumes that there are no false matches across post-strata. Further, all of the models represent a simplification of the underlying sample design of the PES.

3. ESTIMATE THE POPULATION TOTAL

In this section, we consider estimation of the population total under the various matching models, (A), (A′), (B), and (B′), assuming the validity of usual assumptions of independence of the two samples and homogeneous probabilities of inclusion in the samples. For models involving heterogeneous catchability and/or dependence, see the three-sample approach in Darroch *et al.* (1993) and the approach in Alho *et al.* (1993).

Let N be the number of individuals in the population under consideration, x_{1+} the number of individuals in Y_1, x_{+1} the number of individuals in Y_2, and x_{11} the number of individuals in both samples. The number of individuals observed in Y_2 but not Y_1 is $x_{21} = x_{+1} - x_{11}$ and the number observed in Y_1 but not Y_2 is $x_{12} = x_{1+} - x_{11}$.

One can arrange the capture-recapture data in a 2×2 contingency table with one missing cell:

		Sample Y_2 present	Sample Y_2 absent
Sample Y_1	present	x_{11}	x_{12}
	absent	x_{21}	$-$

where we use symbol "$-$" to indicate the missing cell, and standard notation for marginal totals: $x_{1+} = x_{11} + x_{12}$, $x_{+1} = x_{11} + x_{21}$. There is a corresponding 2×2 table of probabilities, $p_{ij} = \Pr[\text{any individual falls into } (i,j) \text{ cell}]$,

		Sample Y_2 present	Sample Y_2 absent
Sample Y_1	present	p_{11}	p_{12}
	absent	p_{21}	p_{22}

with the usual linear constraint

$$\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1.$$

Let n be the number of observed different individuals in the two samples, *i.e.*, $n = x_{11} + x_{12} + x_{21}$. If we assume that the samples are randomly selected with homogeneous selection probabilities, then the numbers of individuals in the four cells have a multinomial distribution

$$(x_{11}, x_{12}, x_{21}, N - n) \sim \text{Mult}(N, p_{11}, p_{12}, p_{21}, p_{22}).$$

We use the conditional likelihood approach developed by Sanathanan (1972). For fixed n, (x_{11}, x_{12}, x_{21}) has a multinomial distribution with likelihood function

$$L_1(p_{11}, p_{12}, p_{21}) = \frac{n!}{x_{11}! x_{12}! x_{21}!} \cdot \frac{p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}}}{(p_{11} + p_{12} + p_{21})^n}. \quad (1)$$

Then n is viewed as being binomially distributed with sample size N and probability $p_{11} + p_{12} + p_{21}$, and the corresponding likelihood is

$$L_2(N) = \frac{N!}{n!(N-n)!} (p_{11} + p_{12} + p_{21})^n [1 - (p_{11} + p_{12} + p_{21})]^{N-n}. \quad (2)$$

In the conditional approach we derive maximum likelihood estimates for the cell probabilities based on the likelihood (1), then find the value of N which maximizes (2), given

the values of the cell probabilities. Sanathanan (1972) has shown that under suitable regularity conditions both conditional and unconditional likelihood estimates of N are consistent and have the same asymptotic multivariate normal distribution. The conditional approach is particularly suitable for a large sample problem like ours.

Under the equal catchability assumption, we let p_1 be the probability that any individual in the population is included in Y_1, and similarly we let p_2 be the probability of inclusion in Y_2. The probabilities p_1 and p_2 are usually referred to as capture probabilities and they do not depend on how the matching mechanism operates. Then the probability that an individual is in both samples is p_1p_2, and the probability of being in set N_1 is $p_1(1 - p_2)$. Since model (A) is a special case of model (B) with $\beta = 0$, we focus on formulating the problem under model (B). To do this, we first need to work out the parametric specification of the cell probabilities. An individual will fall into the (1,1) cell in the 2×2 table only in two cases, *i.e.*, the individual is actually in both samples and a match is made, or, using the notation in the last section, an individual who is actually in N_1 is incorrectly matched with some one in N_2. Here the matching direction from N_1 to N_2 is implicitly assumed in (iii) of model (B). The probability that the former case occurs is αp_1p_2, and the probability that the latter case occurs is $\beta p_1(1 - p_2)$. Furthermore, the two cases are mutually exclusive. Thus, we have $p_{11} = \alpha p_1p_2 + \beta p_1(1 - p_2)$, and, $p_{12} = p_1 - p_{11} = p_1 - \alpha p_1p_2 - \beta p_1(1 - p_2)$, $p_{21} = p_2 - p_{11} = p_2 - \alpha p_1p_2 - \beta p_1(1 - p_2)$. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of parameters in a multinomial distribution. His conditions are satisfied by the parameterization of $\{p_{ij}\}$ here.

For $\alpha = 1, \beta = 0$, this setup reduces to the usual two sample problem and there exist well known solutions in closed form for resulting likelihood equations for the conditional likelihood (1) (*cf.* Bishop *et al.* 1975, chap. 6, p. 232), leading to the usual dual-system estimator, $\hat{N}_{\text{DSE}} = x_{1+}x_{+1}/x_{11}$. Otherwise, the maximum likelihood estimates cannot be written in closed form. Once we have $\hat{p}_1$ and $\hat{p}_2$, however, the conditional maximum likelihood estimates for p_1 and p_2, the conditional maximum likelihood estimate for N can be written as

$$\hat{N} = \frac{n}{\hat{p}_1 + \hat{p}_2 - (\alpha - \beta)\hat{p}_1\hat{p}_2 - \beta\hat{p}_1}, \qquad (3)$$

(*cf.* Chapman 1951). Under model (A′) or (B′), for the i-th stratum, one can use the estimates of the parameters computed under model (A) or (B) for the data of that stratum, and then sum over strata for an estimate of the population total.

4. ESTIMATE MATCHING ERROR RATES BY REMATCH STUDY DATA

In what follows, we give estimates of the matching error rate parameters α and β using the data from the Matching Error Study (rematch study), one of the operations conducted by the Census Bureau in the 1986 Los Angeles test census to evaluate the PES. Briefly, the rematch typically operates for a sample of cases, using more extensive procedures, highly qualified personnel and reinterviews to obtain estimates of the bias associated with the previous matching process. For further details, see Childers, Diffendal, Hogan and Mulry (1989). In their discussion of the Matching Error Study in Los Angeles TARO, Hogan and Wolter (1988) state that "The rematch was done independently of the original match, and the discrepancies between the match and the rematch results are adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results."

The data collected in a rematch study can be displayed as in the following table

Rematch Study Data

		Rematch Classification: Matched	Rematch Classification: Not Matched
Original Classification	Matched	y_{11}	y_{12}
	Not Matched	y_{21}	y_{22}

To estimate α and β, we assume that in the original matching process, errors are made according to model (B) and that errors in the rematch process can be disregarded, *i.e.*, the rematch is assumed to be perfect. It then follows that $y_{11} + y_{21}$ is the true number of matches, and thus is fixed, while y_{11} is a random variable having a binomial distribution, *i.e.*, $y_{11} \sim \mathcal{B}(y_{11} + y_{21}, \alpha)$. Thus the maximum likelihood estimate of α is $\hat{\alpha} = y_{11}/(y_{11} + y_{21})$, and the maximum likelihood estimate of the false nonmatch rate γ is $\hat{\gamma} = 1 - \hat{\alpha} = y_{21}/(y_{11} + y_{21})$. By the same argument, $y_{12} \sim \mathcal{B}(y_{12} + y_{22}, \beta)$, and the maximum likelihood estimate of the false match rate is $\hat{\beta} = y_{12}/(y_{12} + y_{22})$.

We can use the estimates of the matching error rates derived here to analyze the data from the rematch study from the Los Angeles test census. Very often, in addition to estimating the size of a population, it is of interest to estimate the size of a subpopulation such as black, white, or a subpopulation at a certain geographical location. In such case, it is more appropriate to allow for heterogeneity

of matching error rates across various population strata by using estimates of matching error rates for each stratum of interest. Such estimates can be obtained by conducting a rematch study within each stratum and then using the derived estimates. Data for applying model (B′) are available from 1990 Census and are analyzed here.

5. APPLICATIONS

5.1 Application of One Stratum Model to 1986 TARO

Hogan and Wolter (1988) present the rematch data from the 1986 Los Angeles TARO. The rematch results for the P-sample are given in Table 1 in the form of a cross-tabulation of match statuses as assigned from the original TARO match and the rematch. Table 2 presents the two way table of data for the 1986 TARO, with no post-stratification. The estimate of the number missed by both systems, 5,870 is approximately the same order of magnitude as census substitutions 5,259 and erroneous enumerations 6,426 (Hogan and Wolter 1988). Rematch results for the E-sample are presented in Table 3. Let CP, EP be the total correct enumeration and erroneous enumeration by production classification, and let CR, ER be the total correct enumeration and erroneous enumeration by rematch classification, then based on the data in Table 3, Hogan and Wolter (1988) conclude that the original rate of erroneous enumerations (EE), EP/(CP + EP) = 325/(325 + 19,269) = .016 should be increased to about ER/(CR + ER) = 411/(411 + 19,334) = .021.

Table 1

Results of 1986 Los Angeles Test Census Rematch Study: P-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			
	Matched	Not Matched	Un-resolved	Total
Matched	16,623	18	55	16,696
Not matched	88	2,164	56	2,308
Unresolved	17	0	132	149
Total	16,728	2,182	243	19,153

Table 2

Data and Dual-System Estimate for 1986 Los Angeles Test Census. Source: Hogan and Wolter (1988)

		PES		
		Counted	Missed	Total
Correct Census Enumerations*	Counted	298,204	45,463	343,667
	Missed	38,503	5,870	44,373
	Total	336,707	51,333	388,040

*Correct Enumerations = Total Census Enumerations − Substitutions − Erroneous Enumerations.

Table 3

Results of 1986 Los Angeles Test Census Rematch Study: E-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			
	Correct Enumer-ation	Erroneous Enumer-ation	Unresolved	Total
Correct enumeration	19,153	28	88	19,269
Erroneous enumeration	41	283	1	325
Unresolved	140	100	223	463
Total	19,334	411	312	20,057

We now reanalyze the data in Table 2 using model (B), but ignoring the unresolved cases in Table 1 because their classification status are unavailable to us. From the data in Table 1 we estimate $\hat{\gamma} = 1 - \hat{\alpha} = 88/(16{,}623 + 88) = .53\%$, and $\hat{\beta} = 18/(18 + 2{,}164) = .82\%$. In Table 4, we present the estimates and associated standard deviations under model (B) and those from the traditional DSE. The standard deviations are computed using asymptotic normality, for details, see Ding (1990, 1993a, 1993b). The estimated undercount is then defined to be undercount = $(\hat{N} - \text{CEN})/\hat{N} \times 100\%$, and CEN is the total census enumerations, *i.e.*, CEN = Correct Census Enumeration + Substitutions + EE = 343,667 + 5,259 + 6,426 = 355,352. The estimates on the last row of Table 4 indicates that the undercount estimate provided by the DSE should be reduced by 8.42% − 8.05% = .37%. We recall that Hogan and Wolter (1988) argue that the original rate of EE should be increased by 2.1% − 1.6% = .5% as a result of information in the rematch study. This then gives an additional adjustment to the estimated undercount of about .5%. Overall, we estimate that the undercount estimate was biased upward by about .9% (assuming the overlapping is negligible, even though two components are not strictly additive).

Table 4

Comparison of Estimates for 1986 Los Angeles Test Census

Parameter	DSE (SD)	MLE from Model (B) (SD)
p_1	.8856 (5.48×10^{-4})	.8892 (5.51×10^{-4})
p_2	.8677 (5.78×10^{-4})	.8712 (5.86×10^{-4})
N	388,040 (87)	386,470 (79)
Undercount (%)	8.42%	8.05%

Table 5

13 Evaluation Post-strata (EPS) for 1990 PES

1	Northeast, Central City, Minority
2	Northeast, Central City, Nonminority
3	U.S., Noncentral City, Minority
4	Northeast, Noncentral City, Nonminority
5	South, Central City, Minority
6	South, Central City, Nonminority
7	South, Noncentral City, Nonminority
8	Midwest, Central City, Minority
9	Midwest, Central City, Nonminority
10	Midwest, Noncentral City, Nonminority
11	West, Central City, Minority
12	West, Central City, Nonminority
13	West, Noncentral City, Nonminority + Indian

Table 6

Dual System Data for 13 EPS of 1990 PES

EPS	x_{1+} (Census)	x_{+1} (P-sample)	x_{11}
1*	5,966,529	4,656,305.09	4,284,132.78
2	9,235,705	8,685,235.79	8,626,362.34
3*	24,255,611	22,628,349.88	21,068,045.55
4	31,173,378	30,150,266.34	29,966,142.62
5*	9,985,055	8,809,620.02	8,249,407.92
6	13,977,529	13,582,482.34	13,278,614.01
7	47,548,548	44,059,397.93	42,987,517.59
8*	4,060,286	3,714,168.27	3,520,314.04
9	11,826,352	10,058,288.52	9,854,052.95
10	39,343,787	38,358,735.32	38,031,852.01
11*	7,283,885	5,743,998.39	5,365,961.67
12	11,073,872	10,512,339.59	10,222,147.69
13	26,415,232	26,721,116.28	26,025,370.25

*Corresponds to minority post-stratum.

Table 7

Results of Rematch Study for 13 EPS of 1990 PES: P-Sample

EPS	y_{11}	y_{21}	y_{12}	y_{22}
1*	14,301	124	31	2,773
2	15,051	36	16	1,136
3*	28,784	293	49	4,166
4	32,753	703	27	2,058
5*	28,674	189	18	3,738
6	21,757	69	36	1,156
7	48,061	47	20	3,278
8*	14,800	58	21	2,527
9	16,527	39	20	874
10	43,721	120	107	1,664
11*	12,522	133	11	2,097
12	15,122	59	8	1,078
13	43,356	232	108	4,583

Table 8

Results of Rematch Study for 13 EPS of 1990 PES: E-Sample

EPS	CP	EP	CR	ER
1*	17,027	1,415	17,106	1,645
2	15,821	879	15,631	932
3*	32,420	2,430	32,322	2,446
4	33,369	1,242	32,922	1,665
5*	32,412	1,880	33,030	2,044
6	24,392	1,225	24,336	1,284
7	51,107	2,908	50,929	3,047
8*	17,174	1,518	17,133	1,526
9	18,279	648	18,228	656
10	44,450	1,604	44,584	1,631
11*	13,644	985	13,693	909
12	15,647	522	15,590	583
13	49,647	2,062	49,545	2,334

5.2 Application of Multiple Strata Model to 1990 Census

We now analyze stratified data from the evaluation of the PES carried out as part of 1990 decennial census. Hogan (1993) describes operations and results for the 1990 PES, Mulry and Spencer (1991, 1993) present total error analysis, and Davis *et al.* (1991) report on the PES Matching Error Study (MES). The MES was conducted for each of 13 Evaluation Post-strata (EPS) by geographic region and ethnic group. Of the 13 EPS listed in Table 5, five correspond to substantial minority populations (Blacks and Hispanics), *i.e.*, EPS 1, 3, 5, 8 and 11. In Table 6, we present the dual system data for each of the 13 EPS, and we give, in Table 7 and Table 8, relevant rematch data for the P-sample and E-sample. These data are drawn from the final reports on PES evaluation projects P7 and P10 by the Census Bureau (Davis and Biemer 1991a, 1991b). The P-sample for the 1990 PES consisted of about 172,000 housing units (Hogan 1992). The P-sample data are weighted to get estimates of x_{+1} (P-sample total) and x_{11} (total matches) in the usual analysis of the dual system data and the analysis presented here. Nevertheless, the actual unweighted P-sample data can be used to make inference, see Appendix for comparison between estimates from actual P-sample data and estimates from weighted P-sample data.

In Table 9, we give the usual dual system estimates and standard deviations of the capture probabilities (*i.e.*, coverage rate by Census or P-sample) for each of the 13 EPS. Estimates in Table 10 indicate that there is significant variation in matching error rates across the EPS. Among three EPS with $\hat{\gamma}$ larger than .01%, EPS 3 and EPS 11 are minority post-strata. This suggests that the nonmatch rate may be higher for minority post-strata than for the remainder. On the other hand, there is no clear evidence from the estimates of $\hat{\beta}$ that the false match rate is higher

Table 9

Usual Dual System Estimates and Standard Deviations for 13 EPS of 1990 PES

EPS	$\hat{p}_1$ (SD)	$\hat{p}_2$ (SD)	$\hat{N}$ (SD)
1*	0.92007 (12.57 × 10^{-5})	0.71803 (18.42 × 10^{-5})	6,484,855 (470)
2	0.99322 (2.78 × 10^{-5})	0.93402 (8.17 × 10^{-5})	9,298,737 (67)
3*	0.93105 (5.33 × 10^{-5})	0.86858 (6.86 × 10^{-5})	26,051,987 (540)
4	0.99389 (1.42 × 10^{-5})	0.96127 (3.46 × 10^{-5})	31,364,919 (88)
5*	0.93641 (8.22 × 10^{-5})	0.82618 (11.99 × 10^{-5})	10,663,134 (390)
6	0.97763 (4.01 × 10^{-5})	0.95000 (5.83 × 10^{-5})	14,297,391 (131)
7	0.97567 (2.32 × 10^{-5})	0.90408 (4.27 × 10^{-5})	48,734,156 (359)
8*	0.94781 (11.54 × 10^{-5})	0.86701 (16.85 × 10^{-5})	4,283,875 (190)
9	0.97969 (4.45 × 10^{-5})	0.83322 (10.84 × 10^{-5})	12,071,466 (224)
10	0.99148 (1.48 × 10^{-5})	0.96665 (2.86 × 10^{-5})	39,681,946 (108)
11*	0.93419 (10.35 × 10^{-5})	0.73669 (16.32 × 10^{-5})	7,797,041 (443)
12	0.97240 (5.05 × 10^{-5})	0.92309 (8.01 × 10^{-5})	11,388,243 (164)
13	0.97396 (3.08 × 10^{-5})	0.98524 (2.35 × 10^{-5})	27,121,400 (104)

Table 10

Estimates of Matching Error Rates for 13 EPS of 1990 PES

EPS	$\hat{\gamma}$(%)	$\hat{\beta}$(%)
1*	0.009	0.011
2	0.002	0.014
3*	0.010	0.012
4	0.021	0.013
5*	0.007	0.005
6	0.003	0.030
7	0.001	0.006
8*	0.004	0.008
9	0.002	0.022
10	0.003	0.060
11*	0.011	0.005
12	0.004	0.007
13	0.005	0.023

Table 11

MLEs from Model (B′) and Standard Deviations for 13 EPS of 1990 PES

EPS	$\hat{p}_1$ (SD)	$\hat{p}_2$ (SD)	$\hat{N}$ (SD)
1*	0.92406 (12.68 × 10^{-5})	0.72114 (18.79 × 10^{-5})	6,456,833 (446)
2	0.99464 (2.79 × 10^{-5})	0.93536 (8.30 × 10^{-5})	9,285,474 (92)
3*	0.93896 (5.38 × 10^{-5})	0.87597 (7.01 × 10^{-5})	25,832,352 (279)
4	0.99999 (2.65 × 10^{-5})	0.98070 (3.64 × 10^{-5})	30,731,889 (781)
5*	0.94166 (8.28 × 10^{-5})	0.83080 (12.13 × 10^{-5})	10,603,717 (306)
6	0.97922 (4.03 × 10^{-5})	0.95154 (6.03 × 10^{-5})	14,274,182 (64)
7	0.97600 (2.32 × 10^{-5})	0.90438 (4.30 × 10^{-5})	48,717,792 (338)
8*	0.95034 (11.59 × 10^{-5})	0.86933 (17.06 × 10^{-5})	4,272,459 (159)
9	0.97756 (4.47 × 10^{-5})	0.83141 (11.12 × 10^{-5})	12,097,806 (285)
10	0.99217 (1.50 × 10^{-5})	0.96733 (3.06 × 10^{-5})	39,654,306 (90)
11*	0.94239 (10.46 × 10^{-5})	0.74316 (16.58 × 10^{-5})	7,729,158 (359)
12	0.97561 (5.07 × 10^{-5})	0.92614 (8.10 × 10^{-5})	11,350,674 (101)
13	0.97895 (3.10 × 10^{-5})	0.99029 (2.42 × 10^{-5})	26,983,168 (355)

Table 12

Undercount Percentage and Bias Estimates for 13 EPS of 1990 PES

EPS	UC(DSE)	UC(P)	UC(E)	UC(T)	Bias(P)	Bias(E)	Bias(T)
1*	6.40	5.99	5.30	4.89	0.41	1.10	1.51
2	−0.69	−0.83	−1.05	−1.20	0.14	0.36	0.51
3*	5.59	4.79	5.53	4.72	0.80	0.06	0.87
4	−0.11	−2.17	−1.33	−3.39	2.06	1.23	3.29
5*	5.03	4.49	4.68	4.15	0.53	0.35	0.88
6	1.22	1.06	0.99	0.83	0.16	0.23	0.39
7	1.77	1.73	1.50	1.47	0.03	0.26	0.29
8*	3.52	3.26	3.46	3.20	0.26	0.06	0.32
9	1.05	1.26	1.00	1.21	−0.22	0.05	−0.17
10	0.41	0.34	0.36	0.29	0.07	0.05	0.12
11*	5.26	4.43	5.77	4.94	0.83	−0.51	0.32
12	1.89	1.56	1.51	1.19	0.32	0.38	0.70
13	1.79	1.29	1.28	0.78	0.50	0.51	1.01

for minority post-strata, or the other way around. In Table 11, we give maximum likelihood estimates and standard deviations under model (B′). Heterogeneity in the capture probabilities is significant. This heterogeneity together with the variation in the matching error rates suggests that model (B′) is more appropriate than model (B). The asymptotic standard deviations in Table 9 and 11 appear unusually small comparing to the sample size of N. Ding (1993b) shows that this is a typical feature of the dual system problem when the capture probabilities are very high, as it is the case in census application. Despite very narrow confidence intervals, simulation studies in Ding (1993b) show that the asymptotic normal approximation being used is highly accurate in terms of coverage probability.

Table 12 provides estimates of matching bias of various sources in the undercount estimate by the usual DSE. UC(DSE) is the undercount estimate from the DSE defined in the same way as for the 1986 TARO estimate; UC(P) is the undercount estimate computed by MLE from matching error model to adjust for matching bias in P-sample, and Bias(P) = UC(DSE) − UC(P). Again, following Hogan and Wolter (1988), we define the bias in E-sample operation by Bias(E) = ER/(CR + ER) − EP/(CP + EP), and the undercount estimate correcting for E-sample error by UC(E) = UC(DSE) − Bias(E). Finally the total matching bias by both P-sample and E-sample is Bias(T) = Bias(P) + Bias(E), and the undercount estimate correcting for both sources of error is UC(T) = UC(DSE) − Bias(T). Note that it is possible, as observed for EPS 2 and 4 in Table 12, that undercount estimate is negative, thus indicating an overcount instead. This happens when the DSE (or MLE) is less than CEN, the total census enumeration. The dual system data represents "corrected" census counts with erroneous and other incorrect enumerations excluded from CEN.

For each of Bias(P), Bias(E) and Bias(T), a positive estimate indicates a upward bias in the undercount estimate from the DSE by ignoring the corresponding source of error, that is, UC(DSE) should be reduced by the estimated bias to account for that source of error. For each of UC(DSE), UC(P), UC(E) and UC(T), we get significantly higher undercount figures for each of the five minority post-strata, *i.e.*, EPS 1, 3, 5, 8 and 11. For both Bias(P) and Bias(E), all the bias estimates are positive except for Bias(P) for post-stratum 9 and Bias(E) for post-stratum 11. This supports the common belief that there is usually an upward bias attributable to matching errors in the undercount estimate by the DSE, except for some non-minority geographical areas where in fact there is disproportionately large share of erroneous enumerations.

The effects of the two types of matching errors are well understood. False nonmatches results in upward bias and false matches produce downward bias. The nature of the overall matching bias is then dependent upon which type of matching error dominates. By computing undercount estimates for 1980 Census data with selective pair of γ and β, Ding (1990) concludes that due to high capture probabilities in the census application of the capture-recapture technique, the matching bias is dominated by the false nonmatch rate when the false nonmatch rate (γ) and the false match rate (β) are about the same magnitude. This point can be easily confirmed here. EPS 4 has the largest estimate of γ, $\hat{\gamma}$ = .021% and results in the largest Bias(P) = 2.06%. EPS 3 and EPS 4 have about the same estimate of β, $\hat{\beta}$, .012% and .013%, respectively, but EPS 3 has much smaller Bias(P) = .80%, due to smaller estimate of γ, $\hat{\gamma}$ = .010%. About a .01% difference in $\hat{\gamma}$ gives dramatic difference in Bias(P). For matches and nonmatches with complete data, Fay *et al.* (1988, p. 53) state "Because of sometimes difficult nature of the matching work, false nonmatches probably represent a greater concern than false matches". The data analyzed by our methods include both complete data and data produced as a result of the Bureau's imputation procedure. The sensitivity of our estimates to γ lends some support to the statement by Fay *et al.* when both matching for complete data and matching for imputed data are considered together. On the other hand, a downward bias can be observed when $\hat{\beta}$ is much larger than $\hat{\gamma}$. For EPS 9, $\hat{\beta}$ = .022%, about 10 times as large as $\hat{\gamma}$ = .002%. Thus false matches dominate false nonmatches for this stratum, and we see the only negative (downward) bias, Bias(P) = −.22%.

For a specific matching procedure there is an inevitable trade-off between matching errors and unresolved cases. Depending on the extent of unresolved cases and the imputation algorithm used, the resolution process might yield a significant number of false matches. The empirical evidence accumulated by the Bureau of the Census, as we note above, lends some support for the "unbiasedness" of the missing data mechanism used in the imputation process in our example, but further evidence on the issue is desirable.

6. SUMMARY

In this article, we have presented models and methods for the estimation of population total and census undercount that corrects for matching bias of the usual dual-system estimate in the presence of matching errors. Two sources of information are combined in the estimation procedure, the dual-system or capture-recapture census data, and the data from a matching error study (rematch study). The accuracy of our estimates relies on the assumption that the rematch is error free. Matching error rates are likely not to be homogeneous over different population strata. Model (B′) allows for heterogeneity of matching error rates across various population strata but requires stratified rematch data to estimate the error parameters within strata. The methods presented here generalize the standard theoretical framework for the use of maximum likelihood estimation to accommodate matching errors.

We can adjust for erroneous enumerations in the estimate of EE by the use of rematch data for the E-sample. We obtain an overall matching bias in the DSE by adding two bias components from the P-sample and the E-sample. Our analysis of the 1986 Los Angeles test census data indicates that the upward bias of the DSE in the estimate of the census undercount is just under 1%, thereby lending support to the 1% value used by Hogan and Wolter (1988) in their evaluation study. For the analysis on 1990 Census data, the computational results not only agree with understood aspects of matching bias, but also offer findings that were not previously known.

For simplicity, we have assumed that the PES is (allowing for stratification) based on simple random sampling. The models still need to be adapted to account for the complex sampling design actually used (see Hogan 1992, 1993).

It has been known that the perfect matching assumption does not hold in the application of dual system estimation in the U.S. census. The matching problem in the use of the DSE has two components. The first component involves the missing P-sample enumeration status. The second involves errors in classifying P-sample people as enumerated or not. The present paper provides a method to address both components using dual system data adjusted for imputed enumeration probabilities, and can be of possible value in future censuses provided that the models are adapted to handle the complex survey design of the PES. Ding (1993c) develops estimates to directly address the first component by modifying the usual DSE method and describes the relationship between the proposed estimates and those that result from the application of the Census Bureau's imputation scheme for missing P-sample enumeration status (Schenker 1988, Belin *et al.* 1993).

ACKNOWLEDGMENTS

Fienberg's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada to York University, Toronto, Canada. The authors are grateful to Mary Mulry for furnishing data on 1990 Decennial Census, to Joe Sedransk for suggestions, and to Jay Kadane, Larry Wasserman and Mike Meyer for commenting on an earlier version of this work. An Associate Editor and two referees provided comments that have led to a sharpening of the discussion. The basic models in this manuscript were first developed as part of the first author's Ph.D. thesis at Carnegie Mellon University.

APPENDIX

Comparison of Estimates from Weighted and Unweighted P-Sample Data

For simplicity, we assume a weight $k > 1$ for the P-sample and consider the usual dual system estimation problem. Let $\{x_{ij}\}$ be the cell counts in the 2×2 table for weighted P-sample data and census enumerations, $i, j = 1, 2$ and $ij \neq 22$. One could make inference with unweighted P-sample data and census enumerations deflated by a factor of k to get cell counts $\{y_{ij}\}$, $i, j = 1, 2$ and $ij \neq 22$. Then $x_{ij} = ky_{ij}$, $ij \neq 22$, and $x_{1+} = ky_{1+}$, $x_{+1} = ky_{+1}$. Let the usual dual system estimates derived from $\{x_{ij}\}$ be $\hat{p}_1, \hat{p}_2$ and $\hat{N}_w$, and estimates from $\{y_{ij}\}$ be $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_u$. The estimates are (Bishop *et al.* 1975, chap. 6) $\hat{p}_1 = x_{11}/x_{+1} = y_{11}/y_{+1} = \hat{q}_1, \hat{p}_2 = x_{11}/x_{1+} = y_{11}/y_{1+} = \hat{q}_2, \hat{N}_w = x_{1+}x_{+1}/x_{11} = ky_{1+}y_{+1}/y_{11} = k\hat{N}_u$. Thus if one considers the unweighted P-sample data and uses $\hat{N}_* = k\hat{N}_u$ to estimate the population total, then $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_*$ give the same point estimates as $\hat{p}_1$, $\hat{p}_2$ and $\hat{N}_w$ from weighted P-sample data. From the asymptotic normal distribution of the estimates (Ding 1993b), we have $\text{Var}(\hat{N}_w) = k\text{Var}(\hat{N}_u)$, $\text{Var}(\hat{q}_1) = k\text{Var}(\hat{p}_1)$, $\text{Var}(\hat{q}_2) = k\text{Var}(\hat{p}_2)$. Then $\text{Var}(\hat{N}_*) = k\text{Var}(\hat{N}_w)$, and $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_*$ have larger variance than $\hat{p}_1$, $\hat{p}_2$ and $\hat{N}_w$, respectively. To compute estimates with unweighted P-sample data, one needs to know k and $\{y_{ij}\}$. We emphasize that the trivial case of a constant sampling weight for all cases in the same post-stratum is assumed here for simplicity of discussion. However, the real situation can be complex. For example, Blacks may be sampled at a low probability in a White stratum and are then combined with other Blacks sampled with much higher probabilities.

REFERENCES

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.

BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., and ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88, 1149-1166.

BIEMER, P.P. (1988). Modeling matching error and its effect on estimates of census coverage error. *Survey Methodology*, 14, 117-134.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.

CHAPMAN, D.C. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.

CHILDERS, D., DIFFENDAL, G., HOGAN, H., and MULRY, M. (1989). Coverage Evaluation Research: the 1988 Dress Rehearsal. Paper presented to the Census Advisory Committee of the American Statistical Association and the Census Advisory Committee on Population Statistics at the Joint Advisory Committee Meeting, Alexandria, VA.

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 American Statistical Association*, 88, 1137-1148.

DAVIS, M.C., MULRY, M., PARMER, R., and BIEMER, P. (1991). The matching error study for the 1990 Post Enumeration Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 248-253.

DAVIS, M.C., and BIEMER, P. (1991a). Estimates of P-Sample Clerical Matching Error from a Rematching Evaluation. Report on Post-Enumeration Survey Evaluation Project P7, U.S. Department of Commerce, Bureau of the Census.

DAVIS, M.C., and BIEMER, P. (1991b). Measurement of the Census Erroneous Enumerations: Clerical Error Made in the Assignment of Enumeration Status. Report on Post-Enumeration Survey Evaluation Project P10, U.S. Department of Commerce, Bureau of the Census.

DING, Y. (1990). Capture-recapture Census with Uncertain Matching. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y. (1993a). On the asymptotic normality of multinomial population size estimates with application to the backcalculation estimates of AIDS epidemic. To appear in *Biometrika*.

DING, Y. (1993b). On the asymptotic normality of dual system estimates. Unpublished manuscript.

DING, Y. (1993c). Capture-recapture census with probabilistic matching. Submitted for publication.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in central Los Angeles county. *Survey Methodology*, 14, 71-86.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of American Statistical Association*, 80, 98-131.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. U.S. Department of Commerce, Bureau of the Census.

FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. *Encyclopedia of Statistical Sciences, Supplement Volume*, 181-185.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 test census of Tampa, Florida. *Journal of American Statistical Association*, 84, 414-420.

JEFFERSON, T. (1986). Letter to David Humphreys. *The Papers of Thomas Jefferson*, 22, 62.

HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.

HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

MULRY, M.H., and SPENCER, B.D. (1991). Total error in PES estimates of population: the dress rehearsal census of 1988 (with discussion). *Journal of American Statistical Association*, 86, 839-854.

MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of American Statistical Association*, 88, 1080-1091.

RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14, 87-98.

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

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 159–164
Statistics Canada

A Hypothesis Test of Linear Regression Coefficients with Survey Data

PHILLIP S. KOTT[1]

ABSTRACT

This paper discusses testing a single hypothesis about linear regression coefficients based on sample survey data. It suggests that when the design-based linearization variance estimator for a regression coefficient is used it should be adjusted to reduce its slight model bias and that a Satterthwaite-like estimation of its effective degrees of freedom be made. A very important special case of this analysis is its application to domain means.

KEY WORDS: Design-based; Domain mean; Effective degrees of freedom; Model-dependent; Probability order.

1. INTRODUCTION

Most of statistical theory is analytical in nature. One begins with a set of data and a fairly general stochastic model believed to have generated that data. Statistical theory is then invoked to estimate the parameters of the model and to determine the accuracy of those estimates. Ultimately, the original model may be pared down as the result of a series of statistical tests which often take the form of investigations into whether particular parameter values may be reasonably inferred to be zero.

The bulk of survey sampling theory, by contrast, is not analytical but descriptive. There is a finite population of interest. Information about this population can, in principle, be summarized by means of one or more descriptive statistics (for example, the population mean and median). The survey statistician is constrained by time or budgetary considerations to estimate such statistics using only a sample of population units. He (she) often faces a two-fold problem: first a method of sample selection needs to be chosen, then the population statistic(s) needs to be estimated from the sample. Although it is possible to construct a model-dependent statistical theory for these purposes (see, for example, Royall 1970), most survey statisticians invoke a model-free approach known as design-based sampling theory. In this theory, it is not the sample data values that are stochastic (as they are in model-dependent theory) but the sample selection process. Rao and Bellhouse (1989) provides a useful summary of both design-based and model-dependent theory and of attempts to synthesize the two approaches.

The main concern here will be in the testing of a single hypothesis about linear regression parameters. We will assume that the model is correct and that model errors are normally distributed with a possibly complex covariance structure. Unlike Wu *et al.* (1988), we will not explicitly model the error structure (except, perhaps, at a latter stage). Rather, we will focus our attention on a t-statistic calculated using the linearization variance estimator. That this variance estimator has desirable robustness properties from a model-dependent point of view has been demonstrated by Skinner (1989) and Kott (1991).

This paper will provide methods for reducing the model bias of the linearization variance estimator and for determining its effective degrees of freedom. A very important special case of this analysis is its application to the estimated variance of domain means and the difference of such means. Since the analysis in this paper is strictly model-dependent, the terms "bias" and "variance" will refer to model bias and model variance unless otherwise specified.

2. THE MODEL

Suppose we have a population of M elements that can be fit by the linear model:

$$y_M = X_M\beta + \epsilon_M, \tag{1}$$

where y_M is an $M \times 1$ vector of population values for the designated dependent variable;

X_M is an $M \times K$ matrix of population values for the K designated independent variables;

β is a $K \times 1$ vector of regression coefficients; and

ϵ_M is a normally distributed random vector with mean $\mathbf{0}_M$ and variance Σ_M.

A random sample, S, of m distinct elements is drawn from the population. To allow a certain amount of generality in the sampling design, we assume that the population is divided into L strata. From each stratum h, n_h *distinct* clusters of elements are randomly sampled and denoted $u_{h1}, u_{h2}, \ldots, u_{hn_h}$. A random sample of m_{hj} elements is selected from each cluster hj. The clusters are also referred to as primary sampling units. There are $n = \sum n_h$ primary sampling units in the sample.

[1] Phillip S. Kott, National Agricultural Statistics Service, 3201 Old Lee Highway, Fairfax, VA 22030, U.S.A.

Each sampled element has a designation hji, where h is its stratum, hj its primary sampling unit within h, and i the element itself within hj. Let p_{hji} be the probability that element hji is in the sample, and let $w_{hji} = m/(Mp_{hji})$ be the sampling weight of the element. Observe that the sampling weights have been normalized so that if p_{hji} equals the sampling fraction, m/M, then w_{hji} would be unity.

The linear model in (1) also applies to the elements in sample S:

$$y_S = X_S\beta + \epsilon_S,$$

where y_S, for example, is the $m \times 1$ vector of sampled values for the dependent variable. Let $\epsilon_{hj} = (\epsilon_{hj1}, \epsilon_{hj2}, \ldots, \epsilon_{hjm_{hj}})$ be the error vector for the elements in primary sampling unit hj. Now, ϵ_S can be arranged so that the ϵ_{hj} are stacked one on top of the other. Let $\text{Var}(\epsilon_{hj}) = E(\epsilon_{hj}\epsilon_{hj}')$ be denoted by the $m_{hj} \times m_{hj}$ matrix Σ_{hj}, which need not be diagonal. We assume that the ϵ_{hj} are uncorrelated across primary sampling units, so that Σ_S is block diagonal.

The design-based estimator for β is the weighted least squares estimator:

$$b_W = (X_S'WX_S)^{-1}X_S'Wy_S,$$

where W is the $m \times m$ diagonal matrix of sampling weights. The g-th diagonal value of W is the sampling weight associated with the g-th element of the sample. Clearly, b_W is an unbiased estimator of β under the model in (1).

One can simplify the notation for b_W by letting C be the $k \times m$ matrix $(X_S'WX_S)^{-1}X_S'W$, so that $b_W = Cy_S$. Let D_{hj} be a $m \times m$ diagonal matrix with 1's corresponding to the sampled elements of hj and 0's elsewhere. Furthermore, let $C_{hj} = CD_{hj}$. Finally, let $r_S = y_S - X_S b_W$ be the vector of residuals.

The Taylor series or linearization estimator for the mean squared error of b_W (Shah *et al.* 1977) is

$$\textbf{mse} = \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} A_{hj} r_S r_S' A_{hj}', \qquad (2)$$

where $A_{hj} = C_{hj} - n_h^{-1}\sum C_{hg}$, and the summation is over all the primary sampling units in stratum h. The terms "Taylor series" and "linearization" refer to the derivation of mse using design-based sampling theory. Kott (1991) shows that mse is a nearly unbiased estimator of the model *variance* of b_W under reasonable conditions.

It should be noted that in their derivation of mse, Shah *et al.* assumed that the primary sampling units were chosen with replacement. Here, as in Kott (1991), we are assuming that the primary sampling units are distinct which suggests that they were selected with*out* replacement. The reason for this discrepancy is that the assurance of independence among the selected primary sampling units within a stratum in design-based theory and model-dependent theory has almost opposite requirements. The discrepancy goes away, however, if we assume that the primary sampling units were chosen without replacement but that the goal of design-based regression theory is not to estimate a finite population regression parameter but the limit of that parameter as the population (and the number of primary sampling units per stratum) grows arbitrarily large. See Fuller (1975).

If the model in equation (1) holds and $L > 1$, then there is an alternative to **mse** that is also nearly unbiased. It has the same form as equation (2) except that all n sampled primary sampling units are treated as if they came from a single stratum ($L = 1$). Since the alternative can be expressed using equation (2), there is no need to treat it separately in the analysis that follows.

3. A CONVENTIONAL DESIGN-BASED *t*-STATISTIC

The estimator b_W is a K-vector. In this section we will be interested in the t-statistic used to test the univariate hypothesis that $q\beta = \Theta_0$ for some K element row vector $q = (q_1, q_2, \ldots, q_K)$. The most common example of such an hypothesis addresses whether a particular element of $\beta = (\beta_1, \ldots, \beta_K)$, say β_k, is zero. In this example, all of the q_l would be zero except q_k which would be 1; Θ_0 would also be zero.

If the model in (1) and the null hypothesis that $q\beta = \Theta_0$ are true, then

$$\Theta = (qb_W - \Theta_0)/\{q\textbf{Var}(b_W)q'\}^{1/2}$$

would be normally distributed with mean 0 and variance 1. If $\textbf{Var}(b_W)$ were known, the null hypothesis could be tested by comparing the statistic Θ to a standard normal table. Unfortunately, $\textbf{Var}(b_W)$ must be estimated from the sample. Conventional design-based practice is to compare the statistic

$$t = (qb_W - \Theta_0)/(qmseq')^{1/2}, \qquad (3)$$

to a Student's t distribution with $n - L$ or $(n - L - K)$ degrees of freedom (see Shah *et al.* 1977).

The primary goal of this paper is to investigate and then modify the rather *ad hoc* practice described above using the model in equation (1) and our assumptions that Σ_S is block-diagonal. This will be done by investigating $s^2 = qmseq'$ as an estimator for $v^2 = q\textbf{Var}(b_W)q'$. First, s^2 will be adjusted to reduce its bias; then, a better determination of the adjusted estimator's effective degrees of freedom will be established.

4. THE MODEL BIAS OF s^2

The analysis to be conducted is asymptotic. Many of the results rely on the assumption that n, the number of primary sampling units in the sample, is large. (Formally, we should assume that there are infinite sequences of statistics taking on values as n grows arbitrarily large.) If n is large, then so too must be M and m, the number of elements in the population and the sample, respectively. We will assume that $\max\{m_{hj}\}$ is bounded by a finite value, say $\bar{m}_0$. Thus, m is bounded by $\bar{m}_0 n$ and the number of nonzero elements in the block-diagonal matrix Σ_S is bounded by $\bar{m}_0^2 n$.

The number of columns of X_S, K, is assumed to be fixed, but we have some flexibility concerning the number of strata, L. Either L can stay fixed as n grows arbitrarily large with the n_h/n ratios converging to fixed positive limits, or L/n can converge to a fixed positive limit with $\max\{n_h\}$ bounded.

Our concern here is with providing *sufficient* conditions for the subsequent analysis in the text to hold. The random variable ϕ (formally, the infinite random sequence $\{\phi_n\}$) will be said to be of probability order $n^{-\delta}$, *i.e.*, $\phi = O_P(n^{-\delta})$, when $|E(\phi^2)| < B/n^{2\delta}$ for some finite B. Similarly, the random matrix Φ will be said to equal $O_P(n^{-\delta})$ when each element ϕ_{ij} in Φ satisfies $|E(\phi_{ij}^2)| < B/n^{2\delta}$. When ϕ is not random, the P subscript on O is not needed. The same is true for $\boldsymbol{O}$.

The following assumptions are reasonable given the structure that has been laid out:

(1) $\boldsymbol{C} = (X'WX)^{-1}X'W$ exists and is $\boldsymbol{O}(1/n)$, and

(2) $E(\hat{\Sigma}_{hj}) = \Sigma_{hj} + O(1/n)$, where $\hat{\Sigma}_{hj} = r_{hj}r_{hj}'$.

Assumption 1 assures us that $\mathbf{Var}(b_W) = C\Sigma_S C' = \boldsymbol{O}(1/n)$ since there are m elements in the rows of $\boldsymbol{C}$ and no more than $\bar{m}_0^2 n$ non-zero elements in Σ_S.

The variance of qb_W can be rewritten as $v^2 = \sum\sum v_{hj}/n^2$, where $v_{hj} = n^2 g_{hj}\Sigma_S g_{hj}$, $g_{hj} = qCD_{hj}$, and D_{hj} is a diagonal matrix with 1's corresponding to the sampled elements of primary sampling unit hj and 0's elsewhere. Similarly, $s^2 = qmseq'$ can be rewritten as

$$s^2 = \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (g_{hj} - g_h) r_S r_S' (g_{hj} - g_h)' \tag{4}$$

$$= \sum (n_h/[n_h - 1]) \sum [g_{hj}\hat{\Sigma}_S g_{hj}' - 2g_h\hat{\Sigma}_S g_{hj}' + g_h\hat{\Sigma}_S g_h'],$$

where $g_h = \sum g_{hj}/n_h$, the summation is across the j in h, and $\hat{\Sigma}_S = \sum\sum D_{hj} r_S r_S' D_{hj}$.

Both g_{hj} and g_h are $\boldsymbol{O}(1/n)$ because $\boldsymbol{C} = \boldsymbol{O}(1/n)$ and D_{hj} has a bounded number of non-zero values. Thus, $E(g_{hj}\hat{\Sigma}_S g_{hj}') = g_{hj}\Sigma_S g_{hj}' + O(n^{-3})$, $E(g_h\hat{\Sigma}_S g_{hj}') = g_h\Sigma_S g_{hj}' + O(n^{-3})$, and $E(g_h\hat{\Sigma}_S g_h') = g_h\Sigma_S g_h' + O(n^{-3})$. Consequently, $E(s^2 - v^2) = O(n^{-2})$.

Since $r_S = (I_m - XC)\epsilon_S$ and $E(\epsilon_S\epsilon_S') = \Sigma_S$, $E(r_S r_S') = \Sigma_S - XC\Sigma_S - \Sigma_S C'X' + XC\Sigma_S C'X'$. From equation (4), we can see that $E(s^2) = v^2 - R$, where $R = \sum (n_h/[n_h - 1]) \sum (g_{hj} - g_h)Z(g_{hj} - g_h)'$ and $Z = 2XC\Sigma_S - XC\Sigma_S C'X'$. Now $Z = \boldsymbol{O}(1/n)$, because $\boldsymbol{C} = \boldsymbol{O}(1/n)$, X has a fixed number of columns, and the number of non-zero terms in any column of Σ_S is bounded. This implies $R = O(n^{-2})$. Thus, $-R/v^2$, the relative bias of s^2, is $O(1/n)$.

An alternative estimator for v^2 with a reduced relative bias is

$$s_*^2 = s^2/(1 - s^{-2}\hat{R}), \tag{5}$$

where

$$\hat{R} = \left\{ \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (g_{hj} - g_h)\hat{Z}(g_{hj} - g_h)' \right\},$$

and

$$\hat{Z} = 2XC\hat{\Sigma}_S - XC\hat{\Sigma}_S C'X'.$$

In equation (5), $\hat{R}/s^2$ is used to estimate R/v^2. The variance estimator s_*^2 has been proposed here rather than the more obvious $s^2 + \hat{R}$ as *ad hoc* compensation for the slight relative bias of $\hat{R}$ as an estimator of R.

5. THE RELATIVE VARIANCE OF THE VARIANCE ESTIMATOR

Let $e_{hj} = ng_{hj}\epsilon_S$ so that $\mathrm{Var}(e_{hj}) = v_{hj}^2$, and recall that $v^2 = \sum\sum v_{hj}^2/n^2$. If $\hat{e}_{hj} = ng_{hj}r_S$, then the random variable s^2 can be re-written as

$$s^2 = n^{-2} \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (\hat{e}_{hj} - \hat{e}_h)^2$$

$$= n^{-2} \sum (n_h/[n_h - 1]) \{ \sum (e_{hj} - e_h)^2 - (g_{hj} - g_h)A(g_{hj} - g_h)' \},$$

where $A = 2XCe_S e_S' - XCe_S e_S' C'X'$. It is now possible to show that

$$s_*^2 = n^{-2} \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (e_{hj} - e_h)^2 + O_P(n^{-5/2}).$$

Consider a random variable with a χ^2 distribution with F degrees of freedom. Its relative variance is $2/F$. This suggests a Satterthwaite-like determination of the effective degrees of freedom of s_*^2 (see Satterthwaite 1946); namely,

$$F = \frac{(nv)^4}{\sum_{i=1}^{L}\left\{\sum_{j=1}^{n_h} v_{hj}^4 + \sum_{j'\neq j} v_{hj}^2 v_{hj'}^2/(n_h-1)^2\right\}}, \quad (6)$$

which is approximately 2 divided by the relative variance of s_*^2 (since $s_*^2 \approx n^{-2}\sum_h\{\sum_j e_{hj}^2 + \sum_{j'\neq j} e_{hj}e_{hj'}/(n_h-1)\}$).

What is being recommended here is that one tests whether $q\beta = \Theta_0$ by assuming under the null hypothesis that

$$t_* = (qb_W - \Theta_0)/s_*, \quad (7)$$

has a Student's t distribution with F degrees of freedom, where F is determined using equation (6) and making some assumptions about the v_{hj}. Let us call this test the *adjusted t-test*.

6. A SIMPLE EXAMPLE

Consider a simple random sample of n units, n_1 of which are in a subset of the sample denoted by A and n_2 in the complement denoted $\bar{A}$. Let y_i be the observed value for unit i. Suppose the following linear model holds:

$$y_i = d_i\beta_1 + (1-d_i)\beta_2 + \epsilon_i, \quad (8)$$

where $d_i = 1$ is unit i is in set A, and 0 if i is in $\bar{A}$; and the ϵ_i are independent normally distributed random variables.

Assuming homoscedastic errors, both the model-dependent and design-based regression estimator for β_1 is the simple domain mean, $\bar{y}_A = \sum_{i\in A} y_i/n_1$. The linearization estimator for the variance of this estimator is simply $v_L = (n/[n-1])\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^2$. (It should be noted that when a domain mean is viewed as an analytic parameter, its variance requires no finite population correction; see Fuller 1975).

This linearization estimator, v_L, differs from model-dependent variance estimator: $v_M = [\sum_{i\in A}(y_i - \bar{y}_A)^2 + \sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2]/[n_1(n-2)]$. The advantage of v_L is that, unlike v_M, it is asymptotically unbiased under the model even when the ϵ_i are heteroscedastic. This point was noted by Skinner (1989) and Kott (1991). Unfortunately, there still may be considerable bias for finite n. For example, when $n = 100$ and $n_1 = 10$, the relative bias of v_L is approximately 10%. We can see this by noting that $v_E = \sum_{i\in A}(y_i - \bar{y}_A)^2/(n_1[n_1-1]) = ([n-1]/n)(n_1/[n_1-1])v_L$ is exactly unbiased.

Continuing the example: If one were to calculate a t-statistic using conventional design-based practice, he (she) would not only use a biased variance estimator but would also assume that the statistic has 97 or 99 degrees of freedom (100 sampling units minus one strata minus two regressors, were this last subtraction is not always performed). Under ideal conditions (homoscedastic errors within set A), however, the t-statistic calculated using v_E has a Student's t distribution with only 9 degrees of freedom.

Applying equation (5) to the linearization variance estimator, v_L, produces a variance estimator virtually identical to v_E (since $\hat{R} = [v_L/n_1[1 - n_1/n]]$, s_*^2 differs from v_E by only 0.1%). Assuming identically distributed errors within sets A and $\bar{A}$, calculating the effective degrees of freedom, F, with equation (6) yields 9.99. This is almost exactly one degree too many but clearly better than 97 or 99.

A natural hypthesis to test is whether the domain means, β_1 and β_2, in equation (8) are equal. In other words is $\beta_1 - \beta_2 = \Theta_0 = 0$? Assuming that all units have the same variance, the adjusted t statistic is

$$t^* = \frac{\sum_{i\in A} y_i/n_1 - \sum_{i\in\bar{A}} y_i/n_2}{(1 - s^{-2}\hat{R})^{1/2}s},$$

where

$$s^2 = [n/(n-1)][\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^2 + \sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2/n_2^2],$$

and

$$\hat{R} = [n/(n-1)][(\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^3)(1 - n_1/n) + (\sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2/n_2^3)(1 - n_2/n)].$$

To calculate the effective degrees of freedom for $ns^2/(n-1)$ - and thus t_* - using equation (6), note that $L = 1$, and $v_i \propto 1/n_1^2$ for $i \in A$ while $v_i \propto 1/n_2^2$ for $i \in \bar{A}$. As a result,

$$F = \frac{(1/n_1 + 1/n_2)^2}{(1/n_1^3 + 1/n_2^3 + [\{1/n_1 + 1/n_2\}^2 - 1/n_1 - 1/n_2])/n^2},$$

which is 12.3 when $n_1 = 10$ and $n_2 = 90$. The actual degrees of $ns^2/(n-1)$ (*i.e.*, 2 divided by its relative variance) is reasonably close, 11.1 (the relative variance of $ns^2/(n-1)$ is $2[(n_1-1)/n_1^4 + (n_2-1)/n_2^4]$ divided by $[(n_1-1)/n_1^2 + (n_2-1)/n_2^2]^2$).

What this synthetic example principally shows is how misleading conventional design-based practice can be even with an apparently large sample size. The adjusted t-test is clearly a giant step in the right direction.

It is tempting to try to avoid making an assumption about the v_{hj} and to estimate F with

$$f = \frac{(ns_0)^4 - \sum_{}^{L} \sum_{}^{n_h} 2s_{hj}^4/3}{\sum_{i=1}^{L} \left\{ \sum_{j=1}^{n_h} s_{hj}^4/3 + \sum_{j' \neq j} s_{hj}^2 s_{hj'}^2/(n_h - 1)^2 \right\}}, \quad (9)$$

where $s_{hj}^2 = n^2(g_{hj}\, r_S)^2$. Although f is a consistent estimator of F, its use can produce misleading results as we shall see.

Repeated application of equation (9) on 10,000 simulated data sets constructed under the assumption that the ϵ_i in equation (8) are normal, independent, and identically distributed yielded an average f value for the variance of $\bar{y}_A$ of approximately 11.2 with a standard deviation of about 3.5. In addition to its variability, the average f value is greater than F. This is due to the denominator of equation (9) itself being a random variable. It happens that the value of $1/f$ is roughly 0.100 ($\approx 1/9.99$), as expected. Thus, even though the use of f in equation (9) may seem appealing, it is not recommended.

7. ANOTHER EXAMPLE OF A DOMAIN MEAN

Faced with the simple example of the last section, most design-based statisticians would simply treat the units sampled from set A as an independent simple random sample. The linearization and model-dependent variance estimator would then coincide. In practice, however, samples often involve clustering, stratification, and unequal probabilities of selection. When the domain of interest is not a design stratum, it usually becomes impossible to separate out the domain's sampled elements (which need not be primary sampling units) and treat them as an independent random sample.

An example of such a complex sample is the 1985 Continuing Survey of Intakes by Individuals (CSFII). This was a stratified, multistage survey of the dietary intakes of women from 19 to 50 years of age and children from 1 to 5. There were roughly 140 women in the sample who described themselves as black and 1,150 who described themselves as white.

Assuming that a dietary intake value for each individual was independent and identically distributed, values of the relative variance of the linearization variance estimator (R/s^2 from equation (5)) and its effective degrees of freedom (F from equation (6)) were calculated for the two race domains. The relative bias for white women was .003, while the effective degrees of freedom were 48.1. For black women, the relative bias was 0.026, and the effective degrees of freedom 10.1. Thus, even with a fairly large sample size, the effective number of degrees of freedom for black women was relatively small. The conventional determination of degrees of freedom was around 60 (120 PSU's minus 60 design strata).

8. DISCUSSION

As pointed out earlier, the use of design-based techniques can often provide protection when the model in equation (1) fails. Unfortunately, this protection can not be addressed in the strictly model-dependent framework adopted here. It would be unrealistic, however, to expect a conventional design-based t-statistic to behave any better when the model in equation (1) fails than when it holds.

One potential problem of the modified design-based test statistic suggested here occurs when the model in equation (1) does *not* fail: it may not be very powerful. Power can be lost by estimating regression coefficients with sampling weights and by not modelling the error structure directly.

This loss of power is due to the original design-based formulation and not to our modification of it. In fact, s_*^2 is a design consistent estimator of the design mean squared error of b_W whenever s^2 is. This is because $\hat{R}/s^2$ in equation (5) is also $O_P(1/n)$ from a design-based point of view assuming that the first stage of sampling is conducted *with* replacement.

Returning to the simple example of Section 6 can illustrate the issue of power forcefully. The model-dependent and design-based estimates are the same. If all the ϵ_i are assumed to be identically distributed, then the model-dependent variance estimator, v_M, which depends on the assumption of homoscedasticity, is unbiased and has 98 degrees of freedom. The adjusted design based variance estimator is also virtually unbiased, but it has only 9 degrees of freedom.

Often in practice, it will be prudent to sacrifice power for robustness. When that is the case, equation (6) provides an attractive method of measuring how much power may be lost using a modified design-based t-test (equation (7)) when the assumptions of the model are, in fact, correct. Furthermore, the equation lends itself to sensitivity analyses in which the effects of alternative assumptions about the v_{hj} can be evaluated.

ACKNOWLEDGEMENTS

The author would like to thank the staff of the Beltsville Human Nutrition Research Center for its support of this research and an associate editor and his (her) referees for their helpful comments.

REFERENCES

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, C, 37, 117-132.

KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, 107-112.

RAO, J.N.K., and BELLHOUSE, D.R. (1989). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *American Statistical Association Proceedings Sesquincentennial Invited Paper Sessions*, 406-428.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

SHAH, B.V., HOLT, M.M., and FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.

SKINNER, C.J. (1989). Domain means, regression, and multivariate analysis. In *Analysis of Complex Surveys* (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley, 59-88.

WU, C.J.F., HOLT, D., and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 165–169
Statistics Canada

Matrix Masking Methods for Disclosure Limitation in Microdata

LAWRENCE H. COX[1]

ABSTRACT

The statistical literature contains many methods for disclosure limitation in microdata. However, their use by statistical agencies and understanding of their properties and effects has been limited. For purposes of furthering research and use of these methods, and facilitating their evaluation and quality assurance, it would be desirable to formulate them within a single framework. A framework called *matrix masking* - based on ordinary matrix arithmetic - is presented, and explicit matrix mask formulations are given for the principal microdata disclosure limitation methods in current use. This enables improved understanding and implementation of these methods by statistical agencies and other practitioners.

KEY WORDS: Statistical confidentiality; Survey data processing; Mathematical methods.

1. INTRODUCTION

In this Information Age critical activities of society are fuelled by data. Users of statistical data rely especially upon government statistical agencies to collect reliable data and disseminate it in a timely and broadly useful way. Prior to the 1950s, data were released only in printed, tabulated form. Beginning in the 1960s, data at the individual respondent level - *statistical microdata* - began to be released by the U.S. Government.

At present, use of microdata outside statistical agencies for research and policy analysis is often curtailed because appropriate data are not released to users due to confidentiality concerns. For three decades statistical agencies have wrestled with policy and technical issues in microdata release, many of which remain unresolved (Federal Committee on Statistical Methodology 1994). The purpose of this article is to present a class of matrix transformations of microdata intended to help deal with this issue.

Duncan (1990) and Duncan and Pearson (1991) characterized several disclosure limitation methods for microdata - *microdata masks* - by means of matrix addition and multiplication, and named such characterizations "matrix masks." Cox (1991) generalized the concept of matrix masks, and extended the characterization to other microdata masks. The characterization of microdata masks as matrix masks offers conceptual and statistical advantages. Matrix masking provides a simple language to represent, compare and evaluate microdata masking methods. Matrix masking expresses complicated, diverse methods in a form presentable to a wide audience including statisticians and data users, and offers a standard format to develop and optimize the efficiency of transportable microdata masking software.

In this paper, the concept of matrix masks is developed in a mathematically rigorous way. Explicit matrix mask formulations are provided for the principal microdata masking methods in current use, extending those presented in Duncan and Pearson (1991) and Cox (1991). This enables straightforward implementation of these methods in software, and facilitates closer examination and use of microdata masks by statistical agencies. This should lead to improved understanding of the properties of microdata masks and much needed understanding of their effects on data use.

2. MATRIX MASKS

2.1 Definitions

A microdata file containing p attribute values for each of n (respondent-level) data records can be represented as an $n \times p$ matrix X whose entries are denoted x_{ij}. Unless stated otherwise, X contains no missing values. A *matrix mask* (A, B, C) is a transformation of X of the form: $\tilde{X} = AXB + C$, with $A, B \neq 0$, involving ordinary matrix addition and multiplication. As A operates across the rows of X, A is called a *record transforming mask*. B is an *attribute transforming mask*, and C is a *displacing mask* (Duncan and Pearson 1991).

An *elementary matrix mask* of X is a matrix mask of the form AX, XB, or $X + C$. Iterations of (elementary) matrix masks of X are also matrix masks of X. Therefore, a matrix mask of X has the form $\tilde{X} = A\hat{X}B + C$, where either $\hat{X} = X$ or $\hat{X}$ has been obtained from X by application of a sequence of elementary matrix masks. An important advantage of this definition is to enable different statistical disclosure limitation methods to be applied selectively to arbitrary subsets of the records and attributes of X (Section 4).

[1] Lawrence H. Cox, Senior Statistician, U.S. Environmental Protection Agency, AREAL (MD-75), Research Triangle Park, NC 27711, U.S.A.

The matrices A, B, C are not necessarily fixed. For example, a common mask for numeric attributes involves addition of random noise (Tendick 1991), so that C is a random matrix. The matrices A, B, C may depend upon X. For example, to displace X by additive random noise proportional to size, draw the c_{ij} randomly from a normal distribution with mean zero and standard deviation a multiple of $| x_{ij} |$, and set $\tilde{X} = X + C$. Or, with $A = X'$, $M = AX$ is sufficient for ordinary least squares regression (Duncan and Pearson 1991).

2.2 Notation

I denotes the identity matrix. Z denotes the matrix all of whose entries are zero, and J the matrix of all ones. U_{ij} denotes the matrix all of whose entries equal zero, except $u_{ij} = 1$. I is always a square matrix; Z, J and U_{ij} need not be. The U_{ij} matrix, when used as a pre-(post-)multiplier retains the values of only one row (column) of the matrix it multiplies. The dimensions of submatrices may vary between or within individual formulations and will be specified for clarity.

3. REPRESENTATIONS OF DATA MASKS AS ELEMENTARY MATRIX MASKS

3.1 Removing and Selecting Microdata

The most intuitively obvious method for limiting disclosure is to withhold certain microdata from release to data users. Typically, these data are associated with the highest disclosure risk and may require suppressing attributes (columns) or suppressing records (rows) of X prior to release.

Attribute suppression of the k-th attribute can be represented as an attribute transforming mask $\tilde{X} = XB$, where B is the $p \times (p - 1)$ block matrix:

$$\mathbf{Supp}(k) = \begin{bmatrix} I & Z \\ Z & \\ Z & I \end{bmatrix},$$

whose upper I-matrix is of dimension $(k - 1) \times (k - 1)$, whose lower I-matrix is of dimension $(p - k) \times (p - k)$, and whose central Z-matrix is of dimension $1 \times (p - 1)$. An alternative formulation is $\mathrm{Supp}(k) = \sum_{j<k} U_{jj} + \sum_{j>k} U_{j,j-1}$.

Suppression of several attributes can be represented as a product of B-matrices of this form. For example, $\mathbf{Supp}(k)\mathbf{Supp}(j)$ first suppresses the k-th attribute of X, and then suppresses the j-th attribute of the resulting $n \times (p - 1)$ dimensional matrix $X\mathbf{Supp}(k)$. The dimensions of $\mathbf{Supp}(k)$ and $\mathbf{Supp}(j)$ are $p \times (p - 1)$ and $(p - 1) \times (p - 2)$.

It is sometimes necessary to delete individual records from X. For example, a respondent may have high identification risk, or a record may be out of scope or spurious. *Record deletion* of the h-th record can be represented as a record transforming mask $\tilde{X} = AX$, where A is an $(n - 1) \times n$ dimensional block matrix identical in structure to $\mathbf{Supp}(h)$, except: the central Z-matrix of A is of dimension $(n - 1) \times 1$ and the dimensions of the upper and lower I-matrices of A are $(h - 1) \times (h - 1)$ and $(n - h) \times (n - h)$. This A-matrix is denoted $\mathbf{Del}(h)$. An alternative formulation is $\mathbf{Del}(h) = \sum_{i<h} U_{ii} + \sum_{i>h} U_{i-1,i}$.

Deletion of more than one record is represented as a product of A-matrices $\mathbf{Del}(h)$. For example, to delete the h-th and i-th records of X, with $i > h$, use $\mathbf{Del}(i - 1)\mathbf{Del}(h)$. For $i < h$, use $\mathbf{Del}(i)\mathbf{Del}(h)$. The dimensions of $\mathbf{Del}(i - 1)$ and $\mathbf{Del}(h)$ are $(n - 2) \times (n - 1)$ and $(n - 1) \times n$.

The A-matrix that *systematically deletes* every h-th record (for $n = rh$; r an integer) is a block matrix comprising r vertical blocks $\mathbf{Del}(h)$, each of dimension $(h - 1) \times n$. This generalizes to nonsystematic deletion.

The complement of record deletion is *record sampling*. The A-matrix that systematically samples every h-th record of X, for $n = rh$, is an $r \times n$ matrix whose q-th row is the $1 \times n$ dimensional U-matrix $U_{1,qh}$. More generally, to draw a sample of size s comprising the records of X indexed by the set $S = \{s_v : v = 1, \ldots, s\}$, use the A-matrix $\mathbf{Sam}(X, S)$ of dimension $s \times n$, each row of which is a U-matrix U_{1,s_v} of dimension $1 \times n$.

3.2 Aggregating and Grouping Microdata

The risk of a respondent being identified and confidential data disclosed tends to decrease as data are more highly aggregated. *Attribute aggregation* and other microdata masks are based on this principle.

The aggregation mask that replaces the first of two attributes (the j-th attribute) by the sum of the two attributes, and deletes the second attribute (the k-th attribute) from X, for $j < k$, can be represented as an attribute transformation $\tilde{X} = XB$, where B is the $p \times (p - 1)$ dimensional block matrix:

$$\mathbf{Agg}(j,k) = \begin{bmatrix} I & Z \\ & U_{1j} \\ Z & I \end{bmatrix}.$$

The upper I-matrix of $\mathbf{Agg}(j,k)$ is of dimension $(k - 1) \times (k - 1)$, the lower I-matrix is of dimension $(p - k) \times (p - k)$, and the central U-matrix U_{1j} is of dimension $1 \times (p - 1)$. Alternative formulations are

$$\mathbf{Agg}(j,k) = \mathbf{Supp}(k) + U_{kj}, \quad \text{for } j < k, \text{ and}$$

$$\mathbf{Agg}(j,k) = \mathbf{Supp}(k) + U_{k,j-1}, \quad \text{for } j > k.$$

Aggregation-deletion over more than two attributes can be represented as a product of B-matrices of this form. Construct B_1 as above to aggregate the first two attributes to a subtotal, replace the first attribute by the subtotal, and delete the second attribute. Proceed iteratively forming $B_2, \ldots, B_{c-1}$ until all summand attributes have been incorporated into the total and deleted. Then $B = B_1 \cdots B_{c-1}$.

An alternative formulation for aggregation of the j-th and k-th attributes, replacement of the j-th attribute, and deletion of the k-th attribute, is given by the B-matrix product $\mathbf{Add}(j,k)\mathbf{Supp}(k)$. Aggregation and replacement of the j-th attribute without deleting the k-th attribute can be accomplished using the $p \times p$ dimensional B-matrix: $\mathbf{Add}(j,k) = I + U_{kj}$. This generalizes to more summands v by adding more U_{vj}. To create a new totals attribute (attribute $p + 1$) from the j-th and k-th attributes without replacing either attribute, form the $p \times (p + 1)$ dimensional B-matrix $[I \mid U_{j1} + U_{k1}]$, whose I-matrix is of dimension $p \times p$, and whose right-hand submatrix is of dimension $p \times 1$. Aggregating another attribute v amounts to adding additional U_{v1} to the right-hand submatrix.

Grouping categorical data, sometimes referred to as *collapsing categories*, is representable as attribute aggregation. Represent each of the c mutually exclusive categories of a categorical variable by a column of X. The absence (presence) of the corresponding trait is represented in each column by 0 (1). Grouping the c attribute categories to form one combined category is simply aggregation across the c attributes, replacing one attribute by the aggregate and deleting the remaining attributes, using B-matrices in the manner described above.

It is sometimes desirable to aggregate attribute values across microrecords. For example, if microrecords can be grouped according to some notion of "similarity" (*e.g.*, age or profession, or total value of shipments or size of work force for businesses in a particular industry), then an alternative to releasing high risk microrecords is to release a microdata file whose records are *microaggregates* or *microaverages* of subsets of the original records.

Record aggregation can be performed in several ways. A typical case is to replace all summands by the corresponding totals. Assume that the records to be microaggregated are arranged consecutively, and denote the respective sizes of the record groups by $n_1, n_2, \ldots, n_s$, where $n = n_1 + n_2 + \ldots + n_s$. Microaggregation can be accomplished using a diagonal block A-matrix of dimension $n \times n$. The main diagonal of A is comprised of an ordered block of square J-matrices of dimension $n_v \times n_v$, $v = 1, \ldots, s$; the remaining entries of A are zero. Under microaggregation (microaveraging), original values are replaced by microaggregates (microaverages) in each record of the aggregation group. Alternatively, in each group one record may be replaced by the microaggregated record while the other records are deleted. This may be accomplished using J-matrices of dimension $1 \times n_v$, in which case the dimension of A is $s \times n$. To construct microaverages in lieu of microaggregates, each J-matrix is replaced by its corresponding $(1/n_v)J$.

3.3 Scrambling Record Order

A microdata file X being prepared for public use is typically derived from a larger data file (*e.g.*, by sampling) or from a more detailed file (*e.g.*, by removal of directly identifying information such as name, address, and social security number). The larger file is often maintained in a prescribed sort order, such as by geography or social security number, and X is apt to inherit this ordering. To reduce disclosure risk, the order of the microrecords of X must be *scrambled*. Record scrambling can be accomplished using a stochastic A-matrix. Given a reordering of the rows (records) of X (*i.e.*, a permutation P of the row numbers $\{1, \ldots, n\}$), then for $P(i) = h$, set the i-th row of A equal to the U-matrix U_{1h} of dimension $1 \times n$. A is denoted $\mathbf{Reo}(P)$. An alternative formulation is $\mathbf{Reo}(P) = \sum_{i=1}^{n} U_{i,p(i)}$.

3.4 Rounding and Perturbing Microdata

Data rounding is used by statistical agencies for several purposes, including disclosure limitation. Integer variables such as age or years worked, or number of children, presented exactly, could be used in combination with other information to identify respondents (Bethlehem, Keller and Pannekoek 1990). *Conventional rounding* (*e.g.*, base 5, remainders 0, 1, 2 are rounded down; remainders of 3, 4 are rounded up), does not preserve additivity to totals, and *controlled rounding*, designed to preserve additivity to totals in one and two way tabulations, may be preferred (Cox and Ernst 1982). Methods are also available for *unbiased controlled rounding* in one- or two-way tables (Cox 1987).

Data perturbation limits disclosure by introducing slight changes to microdata values. Additive perturbation amounts to adding appropriate perturbation values to original values. Additive perturbation values are often drawn randomly from a distribution with mean zero and variance small relative to that of the data. Nonrandom perturbation is also used.

Rounding and additive perturbation can be represented as displacing masks. For each value x_{ij}, the displacement c_{ij} to x_{ij} is computed according to the rounding or perturbation algorithm, with $c_{ij} = 0$ for those values not subject to change. Then, $\tilde{X} = X + C$ is the matrix of rounded (perturbed) values.

3.5 Attribute Topcoding

Attribute topcoding is a method by which, given a predetermined (large) value T_j of the j-th attribute, all values $x_{ij} > T_j$ are replaced by T_j. Given $x_{ij} = f_{ij} T_j + r_{ij}$,

for f_{ij} the integer quotient, and r_{ij} the remainder, $0 \leq r_{ij} < T_j$, compute $t_{ij} = (\mathbf{Max}\{r_{ij}, (T_j + 1)^{f_{ij}} - 1\})$ $\mathbf{mod}(T_j + 1)$. To topcode X, use the displacing mask $\mathbf{Tco}(X) = (t_{ij} - x_{ij})$.

4. REPRESENTATIONS OF DATA MASKS AS MATRIX MASKS

4.1 Selecting and Modifying Attribute-Record Combinations

The formulations of the preceding section, based on elementary matrix masks, are applied to the entire microdata file X, and do not enable selective masking of arbitrary subsets of records (rows) and/or attributes (columns) of X. The ability to selectively manipulate microdata values within subsets of X (*i.e.*, to apply data masks selectively to submatrices of X) is important for disclosure limitation purposes. This can be accomplished by combining elementary matrix masks that enable *subset selection* along rows and columns, or both, in X with elementary matrix masks as presented previously. This is accomplished in three stages.

At the first stage, apply the ignoring mask $\mathbf{Ign}(Q, R) = AXB$, where A is the $n \times n$ dimensional matrix $A = \sum_{i \in Q} U_{ii}$, and B is the $p \times p$ dimensional matrix $B = \sum_{j \in R} U_{jj}$. A leaves the values in the selected rows Q of X unchanged, and replaces all other values by zeroes; B has similar effect on the columns R. At the second stage, apply the appropriate mask or combination of masks M of Section 3 to $\mathbf{Ign}(Q,R)$ to effect the desired changes, yielding $\tilde{X} = M(\mathbf{Ign}(Q,R))$. As M is designed to change only the selected values, then all ignored values - which $\mathbf{Ign}(Q,R)$ replaced by zero - remain zero after applying M. To preserve the dimensions of $\hat{X}$, deletion operations are modified to replace values to be deleted by zero. Finally, restore the ignored original values of X by means of

$$\tilde{X} = M(\mathbf{Ign}(Q,R)) + X - \mathbf{Ign}(Q,R).$$

4.2 Blurring

When the operation M is microaveraging, the formulation of Section 4.1 provides a matrix mask for the data mask *blurring* of Strudler, Oh and Scheuren (1986).

4.3 Data Swapping

Data swapping is a method whereby selected data values are exchanged between selected sets of records, in a manner that ensures that certain one, two and higher-way tabulations remain unchanged (Dalenius and Reiss 1982). Setting $M = \mathbf{Reo}(P)$, where the swapping rule is given by a permutation P of the affected records, Section 4.1 yields a matrix mask for data swapping.

5. CONCLUDING COMMENTS

A formulation based on matrix algebra for representing the principal statistical disclosure limitation methods for microdata has been developed. Computational issues, such as for large files, are not addressed. However, the partitioning methods of Section 4.1 could be used to reduce effective computational size when working with extremely large files.

Matrix masks offer a comprehensive framework in which statistical agencies can develop, evaluate and use reliable microdata disclosure limitation software. Such software could be shared among agencies. Exploration of the uses of matrix masks by U.S. statistical agencies has been encouraged by an expert panel (Federal Committee on Statistical Methodology 1994, p. 82). The potential effect of the widespread use of matrix masks would be to standardize the microdata disclosure limitation methods available for use by agencies, while expanding each agency's options to evaluate and apply these methods.

ACKNOWLEDGEMENTS

The author is indebted to Professor George T. Duncan, Carnegie Mellon University, for introducing the concept of matrix masks and for collaborations leading to an earlier version of this paper, and to Sumitra Mukherjee, Duncan's doctoral student, for his critical reading and for developing some of the alternative formulations presented here. Preliminary research on this topic was supported in part by National Science Foundation Grant SES 91-10512. The views expressed are those of the author and are not intended to represent the policies or practices of the U.S. Environmental Protection Agency.

REFERENCES

BETHLEHEM, J.G., KELLER, W.J., and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 398, 520-524.

COX, L. (1991). Comment (on Duncan, G.T. and R.W. Pearson 1991, below), *Statistical Science*, 6, 232-234.

COX, L., and ERNST, L. (1982). Controlled rounding. *INFOR*, 20, 423-432.

DALENIUS, T., and REISS, S. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

DUNCAN, G.T. (1990). Inferential disclosure-limited microdata dissemination. *Proceedings of the Survey Research Section, American Statistical Association*, 440-445.

DUNCAN, G.T., and LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7, 207-217.

DUNCAN, G.T., and PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6, 219-239.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and Budget, Washington, DC.

STRUDLER, M., OH, L., and SCHEUREN, F. (1986). Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 375-381.

TENDICK, P. (1991). Optimal noise addition for preserving confidentiality in multivariate data. *Journal of Statistical Planning and Inference*, 27, 341-353.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 171–176
Statistics Canada

Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey

P.D. FALORSI, S. FALORSI and A. RUSSO[1]

ABSTRACT

The study was undertaken to evaluate some alternative small areas estimators to produce level estimates for unplanned domains from the Italian Labour Force Sample Survey. In our study, the small areas are the Health Service Areas, which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut across boundaries of the design strata. We consider the following estimators: post-stratified ratio, synthetic, composite expressed as linear combination of synthetic and of post-stratified ratio, and sample size dependent. For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study in which the sample design was simulated using data from the 1981 Italian Census.

KEY WORDS: Small area estimators; Unplanned domains; Bias; Mean Square Error; Simulation study.

1. INTRODUCTION

In Italy, as in many other countries, there is a growing need for current and reliable data on small areas. This information need concerns most sample surveys realised by the Italian National Statistical Institute (ISTAT), especially the Labour Force Survey (LFS), which has been studied to warrant accuracy in regional estimates.

In the past, ISTAT's solution to this problem was to broaden the sample without changing the estimation method (Fabbris *et al.* 1988). In the last few years, however, in order to find a solution to the negative aspects of oversized samples, research has been launched to identify estimation methods to improve the accuracy of small areas estimates (Falorsi and Russo 1987, 1989, 1990 and 1991).

In our study, the small areas are the Health Service Areas (HSA), which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut-across the boundaries of the design strata. The sizes of these territorial domains are such that the reliability of regular estimates would have been satisfactory had these domains been designed with separate fixed sample sizes from individual domains.

The study was undertaken to evaluate some of the alternative small areas estimators to produce HSA level estimates from the LFS.

We consider the following estimators: post-stratified ratio, synthetic, composite (expressed as linear combination of the synthetic and of the post-stratified ratio), and sample size dependent.

For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study in which the LFS design was simulated using data from the 1981 Italian Census.

2. BRIEF DESCRIPTION OF THE LFS SAMPLE STRATEGY

2.1 Design

The LFS is based on a two stage sample design stratified for the primary sampling units (PSU). The PSUs are the municipalities, while the secondary sampling units (SSU) are the households. In the framework of each geographical region the PSUs are divided according to the provinces. In each province the PSUs are divided into two main area types: the self-representing area consisting of the larger PSUs, and the non self-representing area consisting of the smaller PSUs.

All PSUs in the self-representing area are sampled, while the selection of PSUs in the non self-representing area is carried out within the strata that have approximately equal measures of size. Two sample PSUs are selected from each stratum without replacement and with probability proportional to size (total number of persons). The SSUs are selected without replacement and with equal probabilities from the selected PSUs independently. All members of each sample household are enumerated.

2.2 Estimator of Total

With reference to the generic geographical region, we introduce the following subscripts: h, for stratum ($h = 1, \ldots, H$); i, for primary sampling unit; j, for secondary sampling units; g, for age-sex groups ($g = 1, \ldots, G$).

[1] P.D. Falorsi, Senior Researcher, National Statistical Institute, Rome, Italy; S. Falorsi, Researcher, National Statistical Institute, Rome, Italy; Aldo Russo, Associate Professor, University of Molise, Campobasso, Italy.

In the present study we consider the following age classes 14-19, 20-29, 30-59, 60-64, and over 65.

A quantity referring to stratum h, primary sampling unit i, and secondary sampling unit j will be briefly referred to as the quantity in hij; and a quantity referring to stratum h and primary sampling unit i will be referred to as the quantity in hi.

The following notations are also used: N_h, for number of PSUs in h; P_h, for total number of persons in h; n_h, for number of sample PSUs selected in h; M_{hi} for number of SSUs in hi; P_{hi}, for total number of persons in hi; m_{hi}, for number of sample SSUs selected in hi; P_{ghij}, for number of persons in group g belonging to hij; P_{hij}, for number of persons in hij.

Further let

$$Y = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij}$$

be the total of the characteristic y for regional population, where Y_{ghij} denotes total of the characteristic of interest y for the P_{ghij} persons. Actually, the estimate of Y is obtained by a post-stratified estimator. This estimator is given by:

$$\hat{Y} = \sum_{g=1}^{G} \frac{\hat{Y}_g}{\hat{P}_g} P_g,$$

where

$$\hat{Y}_g = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij}; \quad \hat{P}_g = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij}$$

represent unbiased estimates of

$$Y_g = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij}; \quad P_g = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} P_{ghij}.$$

In the above formulas, the symbol K_{hij}, that denotes the basic weight, is expressed by:

$$K_{hij} = \frac{P_h}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}}.$$

3. SMALL AREA ESTIMATORS

With reference to the generic geographical region, we suppose that the population P is divided into D non-overlapping small areas $1, \ldots, d, \ldots, D$ for which estimates are required. Each area is obtained by an aggregation of municipalities. The problem considered is the estimation the total of a y-variable for all units belonging to the small area d. In practice, the small area d will have a non-null intersection with only a certain number of design strata which we denote as $\tilde{H} = \{h \mid {}_dP_h > 0\}$, where ${}_dP_h$ represents the part of P_h belonging to the small area d.

Denoting by ${}_dN_h$ the number of PSUs belonging to small area d in stratum h, we seek to estimate the small area total

$${}_dY = \sum_{g=1}^{G} \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} Y_{ghij}.$$

The development of a particular estimation method for small areas basically depends on available information. In Italy the accessible information at small area level is very poor. At present the accessible territorial information is total population by sex for each municipality collected through register statistics. In a future context (at end of 1994), the population counts by age-sex group will be available for each municipality. For this reason, in the present study we consider only those small area estimators that utilize, as auxiliary information, the population total by age-sex group.

3.1 Post-stratified Ratio Estimator

A post-stratified ratio estimator (POS) of ${}_dY$ is given by:

$${}_d\hat{Y}_{\mathrm{POS}} = \sum_{g=1}^{G} \frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g} {}_dP_g, \tag{1}$$

where

$${}_d\hat{Y}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij} \delta_{hi},$$

$${}_d\hat{P}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij} \delta_{hi},$$

$${}_dP_g = \sum_{h=1}^{\tilde{H}} {}_dP_{gh} = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} P_{ghij},$$

in which ${}_dP_{gh}$ denotes the total population for the age/sex group g in small area d intersected by stratum h, δ_{hi} is a binary variate that equals 1 if the PSU hi belongs to the small area d and equals 0 otherwise. For a better explanation of formula (1), we observe that PSU is a subset of small area and then does not intersect it.

The post-stratified ratio estimator is unbiased except for the effect of ratio estimation bias which is usually negligible. The estimator is defined to be zero when there is no sample within the domain. This estimator is not reliable for small sample sizes.

3.2 Synthetic Estimator

For computing a synthetic estimator, it is assumed that the small area population means for given population subgroups are approximately equal to the larger area populations means of the same sub-groups. This estimator is obtained by means of a two steps procedure: (i) with respect to an aggregated territorial level, estimates of the investigated features are determined for population subgroups; (ii) estimates for the aggregated territorial level area are then scaled in proportion to the sub-group incidence within the small domain of interest.

The synthetic estimator has a low variance since it is based on a larger sample, but it suffers from bias depending on the distance from the assumption of homogeneity, for each subgroup, between the small area and the larger area with reference to the characteristic of interest, y. The problems associated with synthetic estimators have been documented by Purcell and Linacre (1976), Gonzalez and Hoza (1978), Ghangurde and Singh (1978), Schaible (1979) and Levy (1979) among others.

In this study we consider the following form of synthetic estimator (SYN):

$$ {}_d\hat{Y}_{SYN} = \sum_{g=1}^{G} \frac{\tilde{Y}_g}{\tilde{P}_g}\, {}_dP_g, \tag{2} $$

where

$$ \tilde{Y}_g = \sum_{h=1}^{\bar{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, Y_{ghij};\ \tilde{P}_g = \sum_{h=1}^{\bar{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, P_{ghij}. $$

3.3 Composite Estimator

The composite estimator (COM) considered here is obtained as a linear combination of the estimators SYN (biased with low sample variance) and POS (less biased with high sample variance):

$$ {}_d\hat{Y}_{COM} = \alpha\, {}_d\hat{Y}_{POS} + (1 - \alpha)\, {}_d\hat{Y}_{SYN}, \tag{3} $$

where α is a constant ($0 \leq \alpha \leq 1$). This estimator minimizes the chances of extreme situations (both in terms of bias and sample variance). Therefore, in a given concrete situation such estimator may turn out to be more advantageous than its two components considered separately.

The optimum value for α that minimizes the MSE of the COM estimator is given by

$$ \alpha_{opt} = \frac{\mathrm{MSE}({}_d\hat{Y}_{SYN}) - E({}_d\hat{Y}_{SYN} - {}_dY)({}_d\hat{Y}_{POS} - Y_d)}{\mathrm{MSE}({}_d\hat{Y}_{SYN}) + \mathrm{MSE}({}_d\hat{Y}_{POS}) - 2E({}_d\hat{Y}_{SYN} - {}_dY)({}_d\hat{Y}_{POS} - Y_d)}. \tag{4} $$

Furthermore, when neglecting the covariance term in (4), under the assumption that this term will be small relative to $\mathrm{MSE}({}_d\hat{Y}_{SYN})$ and $\mathrm{MSE}({}_d\hat{Y}_{POS})$, the optimal weight α can be approximated by

$$ \alpha^*_{opt} = \frac{\mathrm{MSE}({}_d\hat{Y}_{SYN})}{\mathrm{MSE}({}_d\hat{Y}_{SYN}) + \mathrm{MSE}({}_d\hat{Y}_{POS})}. \tag{5} $$

This is the approach to define weights followed by Schaible (1978).

In our work the optimal values of α have been obtained from Census data using formula (5). When considering a real sample survey only an estimated value of optimum α may be used, thus resulting in a decrease in efficiency.

3.4 Sample Size Dependent Estimator

The sample size dependent estimator is a particular case of the composite estimator. The linear combination of synthetic and of the less biased estimator is made for each sub-group and depends on the outcome of the given sample. We consider the following form of sample size dependent estimator (SD) which take into account the realized sample size in the small area. It is defined as (Drew, Singh and Choudhry 1982):

$$ {}_d\hat{Y}_{SD} = \sum_{g=1}^{G} \left\{ \alpha_g \left(\frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g}\, {}_dP_g \right) + (1 - \alpha_g)\, \frac{\tilde{Y}_g}{\tilde{P}_g}\, {}_dP_g \right\}, \tag{6} $$

where

$$ \alpha_g = \begin{cases} 1/({}_dR_g F) & 1/{}_dR_g < F, \\ 1 & \text{otherwise} \end{cases} \tag{7} $$

with ${}_dR_g = {}_dP_g / {}_d\hat{P}_g$.

The constant F is chosen to control the contribution of the synthetic component. The reliance on the synthetic portion decreases as the value of F increases. The choice of the value for F would depend upon several factors. In our study the efficiency of sample dependent estimator has been investigated for $F = 1$. This value proved to be efficient while affording protection against the bias of synthetic estimator.

The logic behind the SD estimator is that when the sample size within domain d and group g is small, then the direct estimate for domain d and group g would be unstable and a synthetic estimate may be superior. However, if the sample in domain d and group g is larger than expected this is not a problem, since the performance of the post-stratified direct part would improve as the sample size improves. In conclusion, we observe that SD estimator may be considered as a particular form of sample size dependent regression estimator given in Särndal and Hidiroglou (1989), that has good conditional properties.

4. DESCRIPTION OF THE EMPIRICAL STUDY

4.1 Simulation of the LFS Sample Design

In our study, we have considered the 14 HSAs of the Friuli region as small areas. The variable of interest, y, is the number of unemployed.

Evaluation of the performance of the various estimators, discussed in Section 3, was done by referring to a sample design (two stages with stratification of the PSUs) identical to that adopted for the LFS in Friuli. This design is based on the selection of 39 PSUs and 2,290 SSUs from a population of 219 PSUs and 465,000 SSUs.

We have selected independently 400 Monte Carlo sample replicates each of identical size (in terms of PSUs and of SSUs) of the LFS' sample. All the information utilized in the simulation is taken from the 1981 General Population Census, so ${}_dY$ is known.

4.2 Evaluation of Small Area Estimators

We denote by ${}_d\hat{Y}(mr)$ the estimate of the total ${}_dY$ for the small area d from the rth Monte Carlo replicate when using the estimator m. The percent relative bias of estimator m for the small area d is given by

$$ {}_d\mathrm{ARB}_m = \frac{1}{R}\left(\sum_{r=1}^{R} \frac{{}_d\hat{Y}(mr)}{{}_dY} - 1 \right) 100, $$

where R is the number of samples ($R = 400$).

The average of the percent absolute relative bias of estimator m over the whole set of small areas is:

$$ \overline{\mathrm{ARB}}_m = \frac{1}{D} \sum_{d=1}^{D} | {}_d\mathrm{ARB}_m |, $$

where D is the number of small areas under observation ($D = 14$).

The percent root mean square error of estimator m for small area d is

$$ {}_d\mathrm{RMSE}_m = \frac{\sqrt{{}_d\mathrm{MSE}_m}}{{}_dY} 100, $$

where the mean square error of estimator m for the small area d is expressed by

$$ {}_d\mathrm{MSE}_m = \frac{1}{R} \sum_{r=1}^{R} ({}_d\hat{Y}(mr) - {}_dY)^2. $$

The average percent root mean square error of estimator m over all areas is

$$ \overline{\mathrm{RMSE}}_m = \frac{1}{D} \sum_{d=1}^{D} {}_d\mathrm{RMSE}_m. $$

4.3 Analysis of Results

A. Overall Performance Measures

The average percent absolute biases and the average percent root mean square errors of the small area estimators for the LFS characteristic "number of unemployed persons" are presented in Table 1. Looking at this table, the following conclusions emerge:

(i) As expected, POS presents the smallest bias. The bias of SYN is larger than the bias of the other estimators. The bias of COM is roughly 30% lower than the bias of SYN estimator. The bias of SD estimator is only slightly lower than that of POS estimator.

(ii) SYN and COM have the smallest average percent root mean square errors, but these estimators are affected by a very high bias. POS, with low bias, is, conversely, the less efficient estimator. The average percent root mean square error of SD is approximately 30% higher than those of SYN and COM estimators.

Table 1

Average Percent Absolute Relative Bias $\overline{\mathrm{ARB}}$ and Average Percent Root Mean Square Error $\overline{\mathrm{RMSE}}$ for Unemployed by Estimator

Estimator	$\overline{\mathrm{ARB}}$	$\overline{\mathrm{RMSE}}$
POS	1.75	42.08
SYN	8.97	23.80
COM	6.00	23.57
SD	2.39	31.08

B. Performance Measures by Small Area

Tables 2 and 3 present the Percent Relative Bias (${}_d$ARB) and the Percent Root Mean Square Error (${}_d$RMSE) of the estimators for each of fourteen Health Service Areas in Friuli. Furthermore, Table 2 gives the percent ratio between the population of the HSA and the population of the set $\bar{H}$ of strata including the HSA (p_1); Table 3 shows the percent ratio between the population of the HSA and the population of the region Friuli (p_2) and the percent ratio between the population of the set $\bar{H}$ of strata including the HSA and the population of the region Friuli (p_3). Looking at these Tables, the following conclusions emerge:

(i) SYN and COM are badly biased in some small areas, namely, in those small areas where the model underlying SYN fits poorly. Generally the small areas with low values of the ratio p_1 are affected by large bias (*e.g.*, HSAs 1, 2, 3, 4 and 6). Conversely, large values of the ratio p_1 are associated with low values of the bias (*e.g.*, HSAs 5, 9, 10 and 13). However, SYN and COM consistently have an attractively low RMSE compared to other alternatives. In three of the fourteen areas (*viz*, areas 3, 4 and 8) COM is consistently the most efficient estimator. In two areas (10 and 12)

SYN is evidently more efficient and in the remaining areas the two estimators are roughly similar from the point of view of efficiency. Furthermore, we observe that the lowest values of RMSE for SYN generally are associated with the highest values of the ratio p_3 (*e.g.*, HSAs 1, 2, 5, 6, 9 and 13). HSAs 3 and 4, while having an high value of the ratio p_3, present a high value of RMSE. This is due to the large bias.

(ii) POS shows negligible bias values in almost all small areas. The RMSE values of POS are much higher than those of the other estimators in all the small areas. We observe that the RMSE of the POS estimator is negatively correlated with the ratio p_2. This is caused by the fact that the expected sample size increases as the ratio p_2 increases. Consequently, the variance (which is the main component of MSE of POS) decreases.

(iii) The estimator SD presents a negligible bias in seven (5, 7, 9 10, 11, 12 and 13) of the fourteen small areas. In the other areas the bias is quite low. Furthemore, in nine areas (2, 3, 4, 5, 9, 10, 11, 12 and 13) SD has a bias similar to that of POS. The estimator SD is better, from the MSE point of view, in comparison with POS. In four areas (7, 8, 9, and 13) RMSE is similar to those of SYN and COM.

(iv) Finally, we notice that in the largest areas with the highest values of the ratio p_2 (*e.g.*, HSAs 9 and 5) all the estimators considered give similar results in terms of bias and MSE. For the remaining areas, where the estimators have different performances, there is a problem in the choice of the best estimator.

Table 2

Percent Relative Bias ($_d$ARB) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

HSA	p_1	Estimator			
		POS	SYN	COM	SD
1	19.1	−1.57	−10.92	−7.68	−3.01
2	16.1	−5.61	−9.21	−6.97	−4.79
3	15.3	−5.21	28.82	17.98	5.79
4	16.3	−2.50	20.92	15.02	2.99
5	47.1	−0.46	1.61	0.98	−0.28
6	24.6	−1.37	−12.24	−9.06	−3.28
7	81.8	0.05	−6.25	−3.40	−1.66
8	70.7	0.81	11.80	6.63	2.17
9	92.2	0.47	0.76	0.68	0.78
10	71.2	0.36	−1.34	0.51	−1.02
11	21.7	−1.01	−5.64	−5.00	−1.62
12	40.6	−1.52	−6.66	−6.05	−1.19
13	56.3	−0.95	−3.12	−1.11	−1.28
14	21.8	−2.51	−6.21	−3.03	−3.53

p_1 = percent ratio between the population of the HSA and the population of the set $\bar{H}$ of strata including the HSA.

Table 3

Percent Root Mean Square Error ($_d$RMSE) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

HSA	p_2	p_3	Estimator			
			POS	SYN	COM	SD
1	3.8	19.9	52.23	20.41	21.12	32.39
2	3.1	19.2	63.36	19.45	20.81	38.30
3	3.6	23.2	57.44	36.57	30.71	42.46
4	3.8	23.2	58.19	30.09	27.02	36.88
5	20.2	42.9	18.81	13.38	14.01	17.87
6	8.5	34.8	28.09	17.49	17.00	22.69
7	6.9	8.4	23.83	21.47	21.67	22.67
8	4.8	6.8	28.75	28.54	26.35	27.40
9	21.2	22.9	17.29	16.15	16.40	16.89
10	1.8	2.5	67.00	50.12	53.31	59.27
11	3.2	14.6	49.82	18.35	19.20	30.42
12	4.3	10.7	46.40	22.10	24.04	33.18
13	12.6	22.4	20.13	15.53	15.40	17.88
14	2.3	10.1	57.80	23.58	22.94	36.81

p_2 = percent ratio between the population of the HSA and the population of the region Friuli.

p_3 = percent ratio between the population of the set $\bar{H}$ of strata including the HSA and the population of the region Friuli.

5. CONCLUSIONS

From the point of view of bias, the post-stratified ratio estimator (POS) is essentially unbiased in almost all the small areas. Furthermore the sample size dependent estimator (SD) has negligible values of the bias in almost all small areas. Synthetic (SYN) and composite (COM) estimators present bias values much higher than those of the other estimators.

From the point of view of efficiency, SYN and COM consistently have significantly lower RMSE compared to other alternatives. The estimator SD is much more efficient than POS and furthermore in four of the fourteen areas it shows RMSE values close to those of SYN and COM. Further, when considering the estimator COM there is the problem of the computation of optimum α. In practice only an estimated value of α may be used, resulting in a decrease in efficiency of this estimator. Thus considering both, bias and efficiency, the SD estimator would seem to be preferable to other estimators examined in the context of LFS in Friuli. The sampling rates in Friuli are relatively high and the magnitudes of relative biases and efficiencies of these estimators may be different in other regions where the sampling rates are low, *e.g.*, Piemonte and Lombardia.

REFERENCES

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.

FABBRIS, L., RUSSO, A., and SANETTI, I. (1988). Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 4. Dipartimento di Scienze Statistiche, Universita' di Padova.

FALORSI, P.D., and RUSSO, A. (1987). Un metodo di stima sintetica per piccoli domini territoriali nelle indagini ISTAT sulle famiglie. *Atti del Convegno della Societa' Italiana di Statistica*, Perugia, Italia, 11-20.

FALORSI, P.D., and RUSSO, A. (1989). Un' analisi comparativa di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 18. Dipartimento di Scienze Statistiche, Universita' di Padova.

FALORSI, P.D., and RUSSO, A. (1990). La stima dell'errore quadratico medio di alcune forme di stimatore sintetico nei campioni a due stadi utilizzati nelle indagini ISTAT sulle famiglie. *Giomate di studio: Classificazione ed analisi dei dati, metodi, software, applicazioni*:, Pescara, Italia, 27-39.

FALORSI, P.D., and RUSSO, A. (1991). Evaluation of small area estimation techniques for Italian Labour Force Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 80-106.

GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 52-61.

GONZALEZ, M.E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

LEVY, P.S. (1979). Small area estimation synthetic and other procedures, 1968-1978. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 4-19.

PURCELL, N.J., and LINACRE, S. (1976). Techniques for the Estimation of Small Area Characteristics. Paper presented at the 3rd Australian Statistical Conference, Melbourne.

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

SCHAIBLE, W.L. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 36-83.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 177–184
Statistics Canada

Nonparametric Estimation of Response Probabilities in Sampling Theory

THÉOPHILE NIYONSENGA[1]

ABSTRACT

We deal with the nonresponse problem by drawing on the model of selection in phases that was proposed by Särndal and Swenson (1987). To estimate response probabilities, we use the nonparametric approach first advanced by Giommi (1987). We define estimators according to the nonparametric estimation (NPE) model, and we study their general properties empirically. Inference is based on the concept of quasi-randomization (Oh and Scheuren 1983). The emphasis is on estimating the variance and constructing confidence intervals. We find, by way of a Monte Carlo study, that it is possible to improve the quality of the estimators considered by using a variant of the NPE approach. The latter also serves to confirm the performance of regression estimators in terms of variance estimation.

KEY WORDS: Weighting by phases; Regression estimator; Variance estimators.

1. INTRODUCTION

To counter the effect of nonresponse on the estimation of parameters of a finite population, we consider the phenomenon of nonresponse as a unit selection process in three phases. We therefore use weighting by phases. This adjustment procedure assigns to each unit observed a weight that is inversely proportional to the probability of appearing in the sample, to the unit response probability given the sample, and to the item response probability given the sample and the set of respondents per unit.

In practice, only the probabilities of inclusion in the sample are known. The problem facing us is to estimate individual response probabilities before incorporating them in formulas for the estimators of interest. The nonparametric estimation approach is one of the response probability estimation procedures. It is motivated by the use of auxiliary variables which are linked with unit and item response mechanisms (Giommi 1985, 1987), and which may be correlated with the variables of interest. This avoids assuming that nonresponse is independent of the variables being studied (Oh and Scheuren 1983). This approach also enables us to avoid postulating one or more parametric models governing response, such as the Logit and Tobit models (Grosbras 1987b; Chicoineau, Payen and Thélot 1985) or models of uniform response within subpopulations (Oh and Scheuren 1983; Särndal and Swenson 1985, 1987).

In the Monte Carlo study illustrating certain estimators according to the nonparametric approach, we consider the quite specific case in which the two response mechanisms are governed by the same auxiliary variables. The difference between items will reside in the degree of correlation between each item and the auxiliary variables.

2. NONRESPONSE: A THREE-PHASE SELECTION PROCESS

Consider a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$, of size N. Let s be a sample of fixed size n drawn from U according to a plan $\mathcal{P}(s)$ known and characterized by inclusion probabilities $\pi_k > 0, \forall\, k$ and $\pi_{k\ell} > 0 \,\forall\, k \neq \ell$. We want to observe the units $k \in s$ in relation to a set of Q items $y_1, \ldots, y_q, \ldots, y_Q$ $(Q \geq 1)$, then estimate the total per item $t_q = \sum_U y_{qk}$, for every $q (q = 1, \ldots, Q)$. We assume that conditional on s, each unit k has a probability $\varphi_k > 0$ of participating in the survey and that the probability that two units k and ℓ participate is $\varphi_{k\ell} > 0$ with $\varphi_{kk} = \varphi_k$. We denote the set of units that agree to participate in the survey by r and the mechanism by which the set r was obtained by $\mathcal{P}(r \mid s)$. We further assume that conditional on s and r, each unit $k \in r$ responds to item y_q with probability $\psi_{qk} > 0$ and that the probability that two units k and $\ell \in r$ respond to item y_q is $\psi_{qk\ell} > 0$ with $\psi_{qkk} = \psi_{qk}$. We denote by r_q the set of units that, having agreed to participate in the survey, respond to item y_q and by $\mathcal{P}(r_q \mid s,r)$ the mechanism by which the set r_q is obtained for all $q (q = 1, \ldots, Q)$.

The sets s, r and r_q are obtained from three selection phases for which only the probabilities of inclusion in s are known. The composition of the unit selection mechanisms gives rise to probability outputs that we denote by $\pi_k \Theta_{qk}$ where $\Theta_{qk} = \varphi_k \psi_{qk}$ and $\Theta_{qk\ell} = \varphi_{k\ell} \psi_{qk\ell}$ with $\Theta_{qkk} = \Theta_{qk}$, which do not correspond to inclusion probabilities. Nor does the quantity Θ_{qk} correspond to an inclusion probability for the two response phases conditional on s. If we define the probabilities of inclusion in r_q by $\pi^*_{qk} = \mathbb{P}(k \in r_q)$ and the probabilities of inclusion in r_q given s by $\Theta^*_{qk} = \mathbb{P}(k \in r_q \mid s)$, then (i) $\pi^*_{qk} \neq \pi_k \Theta^*_{qk}$

[1] Théophile Niyonsenga, Ph.D., Researcher, Centre de Recherche Clinique, Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, QC, Canada, J1H 5N4.

and (ii) $\pi_k \Theta^*_{qk} \neq \pi_k \Theta_{qk}$. Furthermore, (iii) $\Theta^*_{qk} = \Theta_{qk}$ if probabilities ψ_{qk} are independent of r, and (iv) $\pi^*_{qk} = \pi_k \Theta_{qk}$ if the φ_k do not depend on s and if the ψ_{qk} do not depend on either r or s.

3. A FEW SPECIAL ESTIMATORS

Assume that there is an auxiliary variable x_q (for the q-th item) strongly correlated with the variable y_q and such that x_{qk} is known $\forall\, k \in s$ or $\forall\, k \in U$. We take the specific case in which $x_{qk} = x_k$, $\forall\, q (q = 1, \ldots, Q)$, and we assume the following linear model ξ

$$\begin{cases} \mathbb{E}_\xi (y_{qk} \mid x_k) = \beta_q x_k \\ \operatorname{Cov}_\xi (y_{qk}, y_{q\ell} \mid x_k, x_\ell) = \begin{cases} \sigma_q^2 x_k & \text{if } k = \ell, \\ 0 & \text{otherwise} \end{cases} \end{cases} \tag{3.1}$$

in which β_q and σ_q are unknown parameters. The following results are extensions of the findings of Särndal and Swenson (1987).

Result 1. If x_k is known, $\forall\, k \in s$, then the regression estimator, denoted by $\hat{t}_{\text{Reg}}$ and defined by:

$$\hat{t}_{\text{Reg}} = \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{x_k}{\pi_k \Theta_{qk}} \right) \sum_s \frac{x_k}{\pi_k}, \tag{3.2}$$

is approximately unbiased for t_q. Its approximate variance is a sum of three components V_1, V_2 and V_3 representing the respective portions of the variance due to the selection phases, that is:

$$V_1 = \sum \sum_U \Delta_{\pi_{k\ell}} (y_{qk}/\pi_k)(y_{q\ell}/\pi_\ell),$$

$$V_2 = \mathbb{E} \left\{ \sum \sum_s \Delta_{\varphi_{k\ell}} (E_{qk}/\pi_k \varphi_k)(E_{q\ell}/\pi_\ell \varphi_\ell) \right\},$$

$$V_3 = \mathbb{E}\,\mathbb{E} \left[\sum \sum_r \Delta_{\psi_{qk\ell}} (E_{qk}/\pi_k \Theta_{qk})(E_{q\ell}/\pi_\ell \Theta_{q\ell}) \mid s \right],$$

where the E_{qk} are theoretical residuals of model (3.1). An estimator of $V(\hat{t}_{\text{Reg}})$ is given by $\hat{V}(\hat{t}_{\text{Reg}}) = \hat{V}_1 + \hat{V}_2^+$ (where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$) with:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{y_{qk}}{\pi_k} \right) \left(\frac{y_{q\ell}}{\pi_\ell} \right), \tag{3.3}$$

and

$$\hat{V}_2^+ = \sum \sum_{r_q} \frac{\Delta_{\Theta_{qk\ell}}}{\Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k \Theta_{qk}} \right) \left(\frac{e_{q\ell}}{\pi_\ell \Theta_{q\ell}} \right), \tag{3.4}$$

where $\Delta_{\pi_{k\ell}} = \pi_{k\ell} - \pi_k \pi_\ell$, $\Delta_{\varphi_{k\ell}} = \varphi_{k\ell} - \varphi_k \pi_\ell$, $\Delta_{\psi_{qk\ell}} = \psi_{qk\ell} - \psi_{qk}\psi_{q\ell}$ and $\Delta_{\Theta_{qk\ell}} = \Theta_{qk\ell} - \Theta_{qk}\Theta_{q\ell}$, the e_{qk} being the observed residuals obtained from model (3.1).

Result 2. If x_k is known, $\forall\, k \in U$, then the regression estimator, denoted by $\hat{t}_{\text{Regl}}$ and defined by:

$$\hat{t}_{\text{Regl}} = N \bar{x}_U \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{x_k}{\pi_k \Theta_{qk}} \right), \tag{3.5}$$

is approximately unbiased for t_q. Its approximate variance is also a sum of three components V_1, V_2 and V_3. The expression of $V_1(\hat{t}_{\text{Regl}})$ differs from that of $V_1(\hat{t}_{\text{Reg}})$ by the use of the theoretical residuals E_{qk} in place of the raw values y_{qk}, whereas the expressions of V_2 and V_3 are identical to those defined above for $\hat{t}_{\text{Reg}}$. An estimator of $V(\hat{t}_{\text{Regl}})$ is given by $\hat{V}(\hat{t}_{\text{Regl}}) = \hat{V}_1 + \hat{V}_2^+$ where:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k} \right) \left(\frac{e_{q\ell}}{\pi_\ell} \right), \tag{3.6}$$

and where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$ is obtained by the formula (3.4).

Comment 1. If $x_k = 1$, $\forall\, k \in U$, the formula (3.5) defines an estimator, denoted by $\hat{t}_{\text{Exp}}$ where:

$$\hat{t}_{\text{Exp}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{1}{\pi_k \Theta_{qk}} = \frac{N}{\tilde{N}} \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}}. \tag{3.7}$$

The estimator $\hat{t}_{\text{Exp}}$ is called an "expansion estimator". An estimator of approximately unbiased variance for $V(\hat{t}_{\text{Exp}})$ is derived from formulas (3.4) and (3.6).

Comment 2. If we take $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall\, k \in U$, in formula (3.7), we obtain an estimator, denoted by $\hat{t}_{\text{Naive}}$, called a "naive estimator". Its expression is given by:

$$\hat{t}_{\text{Naive}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k} \Big/ \sum_{r_q} \frac{1}{\pi_k}. \tag{3.8}$$

If the π_k are constant, the expression (3.8) becomes identical to formula (3.5) in which t is assumed that $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall\, k \in U$, and $x_k = 1$, $\forall\, k \in U$.

Comment 3. For the four estimators defined above, the underlying models are derived from model (3.1) and are the following: $y_{qk} = \beta_q x_k + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2 x_k$ for the first two, $y_{qk} = \beta_q + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2$ and N is known for the last two. For the naive estimator, it is necessary to add the uniform unit and item response model.

4. ESTIMATORS WITH ESTIMATED RESPONSE PROBABILITIES

In practice, the response probabilities φ_k and ψ_{qk} as well as the probability outputs $\Theta_{qk} = \varphi_k\psi_{qk} (k \in U, q = 1, \ldots, Q)$ are actually parameters to be estimated. We estimate them by $\hat{\varphi}_k$, $\hat{\psi}_{qk}$ and $\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\psi}_{qk}$ respectively. We define estimators having the same form as the prototype estimators $\hat{t}_{\text{Exp}}$, $\hat{t}_{\text{Reg}}$ and $\hat{t}_{\text{Regl}}$ seen in section 3, taking care to replace the unknown parameters by their respective estimates. We denote these estimators by $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$ respectively. The variance estimators are obtained from the expressions (3.3), (3.4) and (3.6), in which the unknown parameters are replaced with their estimates.

4.1 Estimation of Response Probabilities

In theory, the probabilities φ_k and ψ_{qk} are functions of the auxiliary variables, that is, functions of the form $\varphi_k = f_1(\upsilon, z_k)$ and $\psi_{qk} = f_2(\mu_q, x_{qk})$ in which the quantities υ and $\mu_q (q = 1, \ldots, Q)$ are unknown parameters and where the pair of vectors (z, x_q), that is, $[(z_1, x_{q1}), \ldots, (z_k, x_{qk}), \ldots, z_N, x_{qN})]'$, contain the auxiliary information available for each item y_q. The nonparametric estimation approach uses only the information contained in (z, x_q) to estimate the φ_k and ψ_{qk}. We are considering here the specific case in which the $z_k = x_{qk} = x_k$, $\forall\, q$ $(q = 1, \ldots, Q)$, and $\forall\, k \in s$.

Let $x_s = \{x_k : k \in s\}$, all the auxiliary information relating to the sample. We specify $\tau_s = \{\tau_k : k \in s\}$, a set of functions such that $\tau_k : \mathbb{R}^n \to \mathbb{R}^1$, for all k in s. We denote by $g_k = \tau_k(x_s)$, $\forall\, k \in s$, the value of the k-th function evaluated in x_s. We subdivide s in n groups s_k not necessarily disjoint, the respective sizes of which are given by:

$$n_k = \sum_{j \in s} D(g_k - g_j), \ (k \in s),$$

$$D(g_k - g_j) = \begin{cases} 1 & \text{if } \mid g_k - g_j \mid \leq h_k, \\ 0 & \text{otherwise,} \end{cases}$$

for a given constant h_k which may depend on all the values g_k $(k \in s)$. The set $s_k = \{j : g_j \in [g_k \pm h_k]\}$, $\forall\, k \in s$, contains j units, whose values g_j vary little from one to another. This group is called the group whose unit k is the kernel, or simply the k-th group. In other words, s_k is a subset of s for which the values of x fall within the vicinity of $x = x_k$ in the sense of the Euclidian distance that specifies $d(k,j) = \mid \tau_k(x_s) - \tau_j(x_s) \mid \leq h_k = h(g_k)$, meaning that $s_k = \{j : d(k,j) \leq h_k\}$. Let $r_k = s_k \cap r$ and $r_{qk} = s_k \cap r_q$. The respective absolute frequencies of these sets are m_k and m_{qk} where:

$$m_k = \sum_{j \in r} D(g_k - g_j), \ (k \in r);$$

$$m_{qk} = \sum_{j \in r_q} D(g_k - g_j), \ (k \in r_q, q = 1, \ldots, Q).$$

Comment 4. In the general case in which nonresponse is governed by the pair of vectors (z, x_q) with $z \neq x_q$, the τ_k functions would be defined in terms of z in order to estimate the unit response probabilities φ_k and in terms of x_q to estimate the item response probabilities ψ_{qk}. Note that this kernel approach can be generalized to more than one auxiliary variable governing response. For two variables x_1 and x_2 governing nonresponse, we would specify the set $s_k = \{(j_1, j_2) : g_{j_1} \in [g_{k_1} \pm h_{k_1}] \text{ and } g_{j_2} \in [g_{k_2} \pm h_{k_2}]\}$.

Response probabilities φ_k and ψ_{qk} are estimated respectively by the rates:

$$\hat{\varphi}_k = \frac{m_k}{n_k}, \ \forall\, k \in r; \ \hat{\psi}_{qk} = \frac{m_{qk}}{m_k}, \ \forall\, k \in r_q, \qquad (4.1)$$

whereas the output $\Theta_{qk} = \varphi_k\psi_{qk}$ is estimated by the rate:

$$\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\varphi}_{qk} = m_{qk}/n_k, \ (k \in r_q, q = 1, \ldots, Q), \qquad (4.2)$$

which is nothing other than the response rate in the k-th group. This simplification of the estimated output $\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\psi}_{qk}$ is, however, possible only when the two response mechanisms are governed by the same auxiliary variables.

Two approaches are considered here: the one based on the values of the variable x (npx) and the one based on the ranks of the values of the variable x (npr). The NPE (npx), proposed by Giommi (1987), is obtained by taking $g_k = \tau_k(x_s) = x_k$ $(k \in s)$. To offset the possible effect of excessively large and excessively small values of x_s, we introduce a variant that consists in using the ranks of x_s, that is, NPE(npr). We consider the function u such that $u(z) = 1$ if $z \geq 0$ and $u(z) = 0$ if $z < 0$. For any unit k in s, let $u_k = \sum_s u(x_k - x_j)$ = the number of components of x_s that are less than or equal to x_k = the rank of x_k in s. The NPE(npr) is then equivalent to letting $g_k = \tau_k(x_s) = u_k (k \in s)$.

4.2 Selection of Interval Limits

The main problem in the NPE approach is the optimum choice of the h_k constants that determine the limits of the intervals $[g_k - h_k; g_k + h_k]$, $\forall\, k \in s$, that is, a choice of $h_k = h_k(g_s)$ that reduces the bias and mean square error of any estimator using the estimated outputs $\hat{\Theta}_{qk}$ specified in formula (4.2).

According to Giommi (1985, 1987), the terms n_k, m_k and m_{qk} that are used to estimate the response probabilities are, apart from the standardization factors, estimators by the kernel method of the density function according to the

approach of Rosenblatt (1956) for the various series of values of g. As an example, it is easy to demonstrate that:

$$n_k = \sum_{j \in s} D(g_k - g_j) = 2nh(n)\hat{f}_n(g_k),$$

where $h(n) = h(g_k, k \in s)$ is a positive constant that converges toward zero at a quite appropriate rate. The theoretical optimum constant, according to the least mean square error criterion, is given by $h(n) = K_f n^{-1/5}$ where K_f, such as defined by Rosenblatt (1956) and Wegman (1972a and b), is obtained by the expression $K_f = [9f(x)/2 \mid f''(x)\mid^2]^{1/5}$.

In practice, $h(n)$ can be obtained only by simulation, since it depends on the density function to be estimated. Giommi (1985) used $h(n) = 2EI_s n^{-1/3}$ where EI_s is the interquartile range in the sample. Kraft, Lepage and van Eeden (1983) chose $h(n) = C(n)EI_s$ where $C(n) = (K_f/EIs)n^{-1/5}$. As our choice, we shall adopt $h(n) = C(n)S_{gs}$, where $C(n) = (K_f/S_{gs})n^{-1/5}$ and where S_{gs} is the corrected standard deviation of the values $g_k(k \in s)$. Basing ourselves on the study of Kraft, Lepage and van Eeden (1983), we will empirically determine a value $\hat{C}_n$ of C that is optimal according to the criterion of least bias and least mean square error of the estimator $\hat{t}^*_{\text{Exp}np}$ and compare the two versions of the NPE approach.

4.3 Expansion and Regression Estimators

Calculation of the approximate bias and variance of the estimators $\hat{t}_{\text{Exp}}$, $\hat{t}_{\text{Reg}}$ and $\hat{t}_{\text{Reg}l}$ is simplified by the fact that the probabilities φ_k and ψ_{qk} are assumed to be known. For estimators $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}lnp}$, these probabilities are estimated by $\hat{\varphi}_k$ and $\hat{\psi}_{qk}$. These probability estimators do not respond to any probability model that would enable us to calculate the bias and the variance conditional on this model. In other words, the sets r_q are generated by unknown response mechanisms for which we estimate the response probabilities by an approach that does not allow for inference conditional on any model underlying the estimation of probabilities.

We would be tempted to resort to Taylor's serial development of the function $1/\hat{\Theta}_{qk}$ to justify the approximation of $1/\hat{\Theta}_{qk}$ by $1/\Theta_{qk}$. In this case, the bias and the variance of $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}lnp}$ would be approached by the approximate bias and variance of $\hat{t}_{\text{Exp}np}$, $\hat{t}_{\text{Reg}np}$ and $\hat{t}_{\text{Reg}lnp}$. However, for sample sizes that are not sufficiently large, we are in danger of having $1/\hat{\Theta}_{qk} \not\equiv 1/\Theta_{qk}$ for the majority of the $k \in r_q$, and consequently:

$$V(\hat{t}^*_{\text{Exp}np}) \not\equiv V(\hat{t}_{\text{Exp}}),\ V(\hat{t}^*_{\text{Reg}np}) \not\equiv V(\hat{t}_{\text{Reg}}),\quad \text{and}$$
$$V(\hat{t}^*_{\text{Reg}lnp}) \not\equiv V(\hat{t}_{\text{Reg}l}).$$

However, to construct confidence intervals based on $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}lnp}$, it is necessary to define estimators for their respective variances. Not having explicit expressions for these variances, it is difficult to define variance estimators and study their properties analytically. The choice of a given estimator is quite difficult to justify. The most natural way of obtaining variance estimators for the variances of $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}lnp}$ is to do a simple substitution of $\Theta_{qk}(=\varphi_k\psi_{qk})$, by $\hat{\Theta}_{qk}(=\hat{\varphi}_k\hat{\psi}_{qk})$, $\forall\, k \in r_q$, and of $\Theta_{qk\ell}$ by $\hat{\Theta}_{qk\ell}$, $\forall\, k \neq \ell \in r_q$ $(\hat{\Theta}_{qk\ell} = \hat{\varphi}_{k\ell}\,\hat{\psi}_{qk\ell})$, in all the formulas for variance estimators specified for the respective variance estimators of estimators $\hat{t}_{\text{Exp}np}$, $\hat{t}_{\text{Reg}np}$ and $\hat{t}_{\text{Reg}lnp}$.

5. MONTE CARLO STUDY: COMPARISON OF ESTIMATORS

For simulation purposes, we assume that Bernoulli trials govern each of the response mechanisms (total or partial) and that a simple random sampling without replacement is the sample design used. We consider a vector $(y_1, y_2, y_3)'$ of three items $(Q = 3)$ and a variable x containing the auxiliary information. We first generate the $x_k(k \in U)$ by a gamma distribution with parameters a_1 and a_2. The generation of items y_1, y_2, y_3 is based on the linear model (3.1) and the gamma distribution. More specifically, we generate the $y_{qk}(k \in U$ and $q = 1, 2, 3)$ according to a gamma distribution with parameters $a_{1q}(x_k)$ and $a_{2q}(x_k)$ defined by:

$$a_{1q}(x_k) = \frac{\beta_q^2 x_k}{\sigma_q^2},\quad a_{2q}(x_k) = \frac{\sigma_q^2}{\beta_q},$$

$$\sigma_q^2 = \beta_q^2 a_2\left\{\frac{1}{\rho^2_{xy_q}} - 1\right\},\quad q = 1, 2, 3.$$

The choice of the gamma distribution is based on its general form, which gives rise to a great variety of distributions, and on the fact that it can represent the distribution of various types of populations (Johnson and Kotz 1970, p. 172). We establish *a priori* the parameters a_1, a_2, β_q and ρ_{xy_q} $(q = 1,2,3)$, namely:

$$a_1 = 2,\quad a_2 = 10,\quad (\beta_1\beta_2\beta_3)' = (0.75\ 0.65\ 0.60)',$$
$$(\rho_{xy_1}\rho_{xy_2}\rho_{xy_3})' = (0.90\ 0.85\ 0.70)'.$$

To generate the unit and item response probabilities, we consider the following exponential forms:

$$\varphi_k = \exp\{-(\lambda_1 x_k + \lambda_2 v_k)\}\quad \text{and}$$
$$\psi_{qk} = \exp\{-(\lambda_{1q} x_k + \lambda_{2q} v_{qk})\},$$

where the v_k and the v_{qk} result from a uniform distribution $(0;1)$. The constants λ_1, λ_2, λ_{1q} and λ_{2q} are such that: $\lambda_1 = 0.15/\bar{x}_U$, $\lambda_{1q} = 0.15/\beta_q\bar{x}_U$ and $\lambda_2 = \lambda_{2q} = 0.45$ $(q = 1,2,3)$. Such a parameterization makes it possible to have an average response rate (total or partial) of approximately 70%. We could have varied these constants or used other continuous functions.

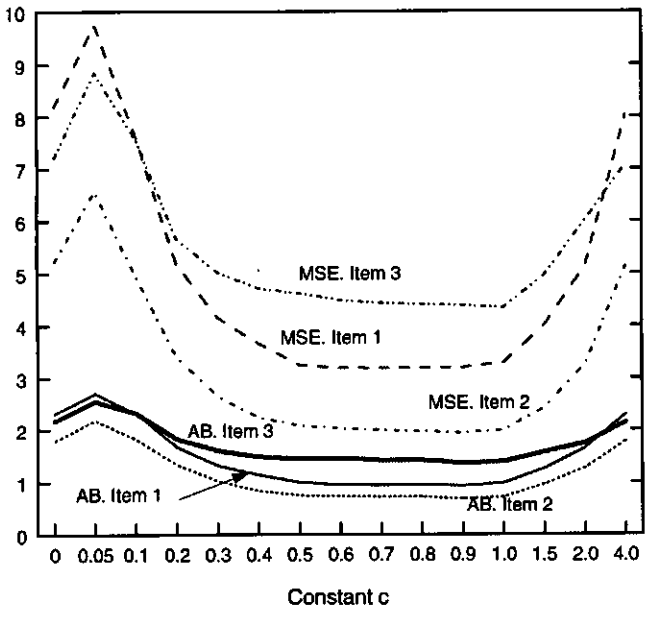

Figure 5.1 Absolute bias and MSE: the estimator $\hat{t}^*_{\text{Exp}npx}$ for $n = 60$

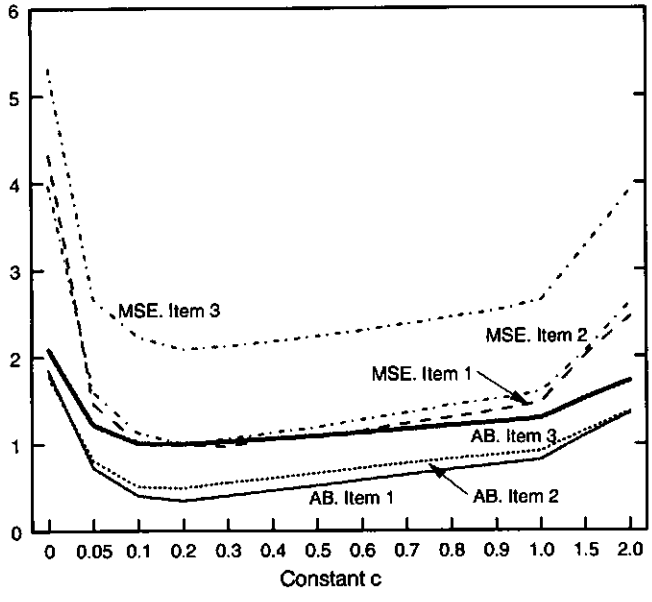

Figure 5.2 Absolute bias and MSE: the estimator $\hat{t}^*_{\text{Exp}npr}$ for $n = 200$

5.1 Comparison of the Two Variants of the NPE Approach

We consider a population of size $N = 100$ and draw a sample s of size $n = 60$, which we subject to the response mechanisms. We repeat the sampling $\mathbb{K}$ times and calculate the bias $\mathbb{B}(\hat{t}^*_{\text{Exp}np})$ and the mean-square error $\text{MSE}(\hat{t}^*_{\text{Exp}np})$, for different values of $C (C \geq 0)$. Next we repeat this experiment with $N = 1{,}000$ and $n = 200$. The results of this empirical study are illustrated by the diagrams of $\mathbb{B}(\hat{t}^*_{\text{Exp}np})$ et $\text{MSE}(\hat{t}^*_{\text{Exp}np})$ as a function of the constant C. From this brief study we observe, firstly, that the value $\hat{C}_n$ of the optimal constant C is in the interval $[0\,;1]$, depends on the size of the sample and decreases as the sample size increases (Figures 5.1 and 5.2).

We also observe that the estimator $\hat{t}^*_{\text{Exp}npr}$ is still better in terms of less bias and mean square error than the estimator $\hat{t}^*_{\text{Exp}npx}$ in the interval $[0;1]$ as illustrated as an example in Figure 5.3 for item 3, the item the least correlated with the auxiliary variable. A very important fact to be noted is that for the estimator $\hat{t}^*_{\text{Exp}npr}$ we more quickly reach the values of the bias and the mean square error of the estimator $\hat{t}_{\text{Naive}}$ in $[0;1]$ at $C = 0.05$ and outside this interval at $C = 4$. Unlike with the estimator $\hat{t}^*_{\text{Exp}npr}$, the values of the bias and the mean square error of the estimator $\hat{t}^*_{\text{Exp}npx}$ first reach maximum values at $C = 0.05$ before taking on the values of the bias and mean-square error of $\hat{t}_{\text{Naive}}$ at $C = 0$. We also note that for a fairly large size n and for any value of C in the interval $[0;1]$, the variation is hardly perceptible (Figure 5.3). For this reason, we suggest that a compromise value be used: $C = 0.5$ (that is, $h = 0.5S_{gs}$).

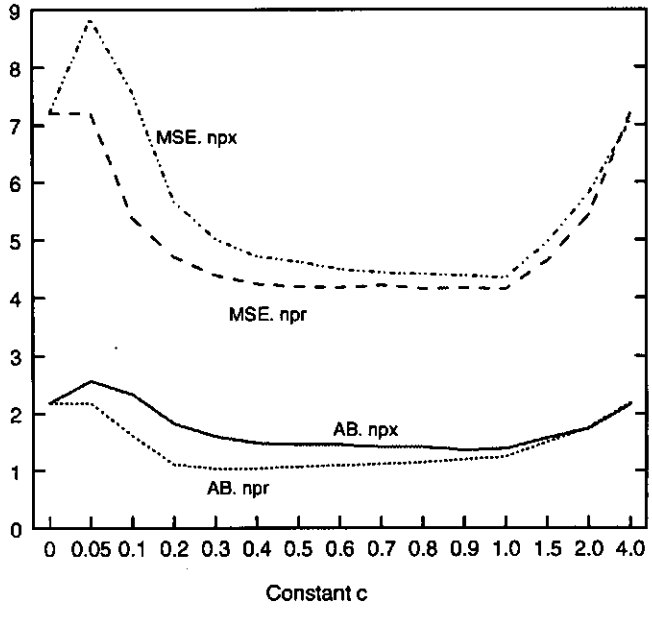

Figure 5.3 Absolute bias and MSE: the estimators $\hat{t}^*_{\text{Exp}npx}$ and $\hat{t}^*_{\text{Exp}npr}$ for item 3

5.2 Overall Comparison of Estimators

The complete operation of the simulation consists in (i) first, drawing the sample s of size $n = 200$ of the population of size $N = 1{,}000$, (ii) then applying the unit and item response mechanisms to obtain sets $r_q (q = 1,2,3)$, and (iii) lastly, calculating, for each estimator, the values

of $\hat{t}$ and $\hat{V}(\hat{t})$. We repeat this operation $\mathbb{K}$ times. Once the experiment is completed, we calculate, as performance measurements, (i) the bias $\mathbb{B}(\hat{t}) = \mathbb{E}(\hat{t}) - t_q$, (ii) the mean square error $\mathrm{MSE}(\hat{t}) = \mathbb{E}(\hat{t} - t_q)^2$, (iii) the expectation of the variance estimator $\mathbb{E}(\hat{V}(\hat{t}))$ and (iv) the theoretical recovery rate $P_o(\hat{t}) = \mathbb{P}\{\mid \hat{t} - t_q \mid \leq Z_{\alpha/2}[V(\hat{t})]^{1/2}\}$. We can also calculate, for each given estimator, (v) the relative error $\mathrm{RE}(\hat{t})\,[= \mathbb{B}(\hat{t})/t]$, (vi) the variance $V(\hat{t})\,[= \mathrm{MSE}(\hat{t}) - (\mathbb{B}(\hat{t}))^2]$, (vii) the relative bias $\mathrm{RB}(\hat{t})\,[= \mid \mathbb{B}(\hat{t}) \mid /(V(\hat{t}))^{1/2}]$ as well as (viii) the relative error of the variance estimator $\mathrm{RE}(\hat{V}(\hat{t}))$ $[= \mathbb{B}(\hat{V}(\hat{t}))/V(\hat{t})]$ in order to examine the sensitivity of the variance estimators to nonresponse.

5.3 Interpretation of the Results of the Global Simulation

I. The Prototype Estimators

The simulation results confirm the theory. For these estimators, we make the following observations, based on Tables 5.1 to 5.4:

(i) $\hat{t}_{\mathrm{Exp}}$, $\hat{t}_{\mathrm{Reg}}$ and $\hat{t}_{\mathrm{Regl}}$ are approximately unbiased;

(ii) $\mathrm{MSE}(\hat{t}_{\mathrm{Regl}}) < \mathrm{MSE}(\hat{t}_{\mathrm{Reg}}) < \mathrm{MSE}(\hat{t}_{\mathrm{Exp}})$;

(iii) $V(\hat{t}_{\mathrm{Regl}}) < V(\hat{t}_{\mathrm{Reg}}) < V(\hat{t}_{\mathrm{Exp}})$ and $\mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Regl}})] < \mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Reg}})] < \mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Exp}})]$.

For these estimators, we also expected that:

(i) $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Exp}}) \approx V(\hat{t}_{\mathrm{Exp}})$, $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Reg}}) \approx V(\hat{t}_{\mathrm{Reg}})$ and $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Regl}}) \approx V(\hat{t}_{\mathrm{Regl}})$;

(ii) Negligible relative bias $[\mathrm{RB}(\hat{t}) < 0.10]$; the recovery rates are close to the theoretical rates. The relative errors $\mathrm{RE}(\hat{t})$ and $\mathrm{RE}(\hat{V}(\hat{t}))$ are negligible, and are in part due to the simulation (errors due to the limited number of repetitions of the experiment).

Table 5.1
The Values of $\mathbb{B}(\hat{t})$, $\mathrm{MSE}(\hat{t})$

	y_1		y_2		y_3	
$\hat{t}_{\mathrm{Exp}}$	−0.036	1.690	−0.052	1.525	−0.056	2.299
$\hat{t}_{\mathrm{Reg}}$	−0.020	0.735	−0.019	0.744	−0.030	1.446
$\hat{t}_{\mathrm{Regl}}$	−0.012	0.319	−0.012	0.431	−0.021	1.202
$\hat{t}_{\mathrm{Naive}}$	−2.037	5.069	−1.937	4.535	−2.220	5.911
$\hat{t}^*_{\mathrm{Exp}npx}$	−0.690	1.345	−0.777	1.407	−1.228	2.604
$\hat{t}^*_{\mathrm{Exp}npr}$	−0.601	1.175	−0.709	1.249	−1.140	2.345
$\hat{t}^*_{\mathrm{Reg}npr}$	−0.293	0.785	−0.414	0.830	−0.895	1.834
$\hat{t}^*_{\mathrm{Regl}npr}$	−0.285	0.376	−0.407	0.520	−0.886	1.621

Table 5.2
The Values of $V(\hat{t})$, $\mathbb{E}[\hat{V}(\hat{t})]$ and $100*\mathbb{E}[\hat{V}_1(\hat{t})]/\mathbb{E}[\hat{V}(\hat{t})]$

	y_1			y_2			y_3		
$\hat{t}_{\mathrm{Exp}}$	1.689	1.683	29.8	1.525	1.485	29.1	2.296	2.235	26.9
$\hat{t}_{\mathrm{Reg}}$	0.734	0.697	72.2	0.744	0.702	61.5	1.445	1.391	42.6
$\hat{t}_{\mathrm{Regl}}$	0.319	0.293	34.0	0.431	0.402	32.7	1.201	1.130	29.3
$\hat{t}_{\mathrm{Naive}}$	0.918	0.911	43.3	0.784	0.766	43.5	0.983	0.958	44.2
$\hat{t}^*_{\mathrm{Exp}npx}$	0.869	1.403	32.0	0.804	1.173	32.3	1.097	1.322	35.4
$\hat{t}^*_{\mathrm{Exp}npr}$	0.814	1.291	35.1	0.746	1.089	35.2	1.046	1.285	37.1
$\hat{t}^*_{\mathrm{Reg}npr}$	0.700	0.627	73.9	0.658	0.588	66.6	1.033	0.955	50.5
$\hat{t}^*_{\mathrm{Regl}npr}$	0.294	0.259	36.7	0.355	0.315	37.6	0.836	0.751	37.1

Table 5.3
The Values of $\mathrm{RE}(\hat{t})$ and $\mathrm{RE}(\hat{V}(\hat{t}))$

	y_1		y_2		y_3	
$\hat{t}_{\mathrm{Exp}}$	−0.0024	−0.0015	−0.0040	−0.0242	−0.0045	−0.0267
$\hat{t}_{\mathrm{Reg}}$	−0.0014	−0.0510	−0.0015	−0.0556	−0.0024	−0.0373
$\hat{t}_{\mathrm{Regl}}$	−0.0008	−0.0812	−0.0009	−0.0684	−0.0017	−0.0596
$\hat{t}_{\mathrm{Naive}}$	−0.1377	−0.0083	−0.1474	−0.0230	−0.1787	−0.0260
$\hat{t}^*_{\mathrm{Exp}npx}$	−0.0466	0.6141	−0.0591	0.4582	−0.0988	0.2046
$\hat{t}^*_{\mathrm{Exp}npr}$	−0.0406	0.5860	−0.0540	0.4591	−0.0917	0.2282
$\hat{t}^*_{\mathrm{Reg}npr}$	−0.0198	−0.1038	−0.0315	−0.1077	−0.0720	−0.0752
$\hat{t}^*_{\mathrm{Regl}npr}$	−0.0193	−0.1191	−0.0310	−0.1124	−0.0713	−0.1015

Table 5.4
The Levels $P_o(\hat{t})$ at 90%, 95% and the $\mathrm{RB}(\hat{t})$

	y_1			y_2			y_3		
$\hat{t}_{\mathrm{Exp}}$	0.873	0.922	0.027	0.870	0.914	0.042	0.852	0.904	0.037
$\hat{t}_{\mathrm{Reg}}$	0.881	0.929	0.024	0.876	0.929	0.022	0.870	0.917	0.025
$\hat{t}_{\mathrm{Regl}}$	0.866	0.926	0.021	0.873	0.923	0.018	0.860	0.914	0.019
$\hat{t}_{\mathrm{Naive}}$	0.322	0.427	2.126	0.298	0.405	2.187	0.287	0.389	2.239
$\hat{t}^*_{\mathrm{Exp}npx}$	0.851	0.906	0.740	0.800	0.874	0.866	0.667	0.758	1.172
$\hat{t}^*_{\mathrm{Exp}npr}$	0.872	0.925	0.666	0.830	0.893	0.820	0.700	0.789	1.114
$\hat{t}^*_{\mathrm{Reg}npr}$	0.839	0.908	0.350	0.806	0.878	0.510	0.712	0.789	0.880
$\hat{t}^*_{\mathrm{Regl}npr}$	0.804	0.871	0.526	0.767	0.844	0.683	0.678	0.763	0.969

II. The Naive Estimator

The naive estimator registers absolute values of $\mathbb{B}(\hat{t})$ and $\mathrm{RE}(\hat{t})$ that are very high in relation to the other estimators (Tables 5.1 and 5.3). The same is true for the values of $\mathrm{MSE}(\hat{t})$ (Table 5.1). The values of the observed recovery rates $P_o(\hat{t})$ as well as those of the relative bias $\mathrm{RB}(\hat{t})$ are hardly surprising, considering the size of the point estimate bias (Table 5.4).

The behaviour, in terms of variance and variance estimator (Table 5.2) of $\hat{t}_{\text{Naive}}$, is due to the fact that it constitutes a particular case of $\hat{t}_{\text{Exp}}$, assuming uniform response mechanisms. In a sense, this amounts to assuming that the data are missing randomly.

III. The Adjusted Estimators

The reduction of the bias and the mean square error resulting from the use of the adjusted estimators (Table 5.1) is quite significant, in comparison with the naive estimator, especially for the regression estimators (the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$). In terms of variance (Table 5.2), we have the following inequalities:

$$V(\hat{t}^*_{\text{Reg}1npr}) < V(\hat{t}^*_{\text{Reg}npr}) < V(\hat{t}^*_{\text{Exp}npr}) < V(\hat{t}^*_{\text{Exp}npx}),$$

which are analytically difficult to demonstrate. Little variation [in terms of $V(\hat{t})$ and $\mathbb{E}(\hat{V}(\hat{t}))$] is observed between items y_1 and y_2 in light of the little variation between the correlations (0.05). On the other hand, the effect of the correlation with the auxiliary variable on $V(\hat{t})$ and of $\mathbb{E}(\hat{V}(\hat{t}))$ may be observed by comparing items y_1 and y_3, then y_2 and y_3: the variations between the correlations are greater in these two cases (0.20 and 0.15 respectively).

In terms of variance estimators (Table 5.2), we observe that:

$$\hat{V}(\hat{t}^*_{\text{Reg}1np}) < \hat{V}(\hat{t}^*_{\text{Reg}np}) < \hat{V}(\hat{t}^*_{\text{Exp}np}),$$

as such is the case for the estimators $\hat{t}_{\text{Reg}}$, $\hat{t}_{\text{Reg}1}$ and $\hat{t}_{\text{Exp}}$. What is surprising, and is of course due to the effect of the auxiliary variables on the variance components relative to the response mechanisms, is the fact that the estimators $\hat{t}^*_{\text{Exp}np}$ overestimate the variance with very large absolute values of $\text{RE}(\hat{V}(\hat{t}))$, while the regression estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$ underestimate the variance with absolute values of $\text{RE}(\hat{V}(\hat{t}))$ that are smaller in relation to those of $\hat{t}^*_{\text{Exp}np}$ (Table 5.3). For the estimators $\hat{t}^*_{\text{Exp}np}$, not only is the total variance high in relation to that of the regression estimators, but also the relative contribution of the sampling variance is low (Table 5.2).

In terms of recovery rate (Table 5.4), the estimators $\hat{t}^*_{\text{Exp}np}$ yield observed rates that are closer to theoretical rates than the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$. However, the values of the relative bias $\text{RB}(\hat{t})$ are higher for $\hat{t}^*_{\text{Exp}np}$ than for $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$, which makes the confidence intervals less reliable.

IN CONCLUSION

(i) If the goal of the estimation is to reduce bias and mean square error, all the estimators adjusted for nonresponse perform well in relation to the uniform response mechanism (which basically amounts to doing nothing about nonresponse). The rate of reduction of the bias of each estimator in relation to the naive estimator is at least 66%. The regression estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$ are the most promising of the various estimators considered (Table 5.1).

(ii) If the goal is to construct confidence intervals, we need a pair of estimators $[\hat{t}, \hat{V}(\hat{t})]$ that simultaneously minimize the absolute biases $|\ \mathbb{B}(\hat{t})\ |$ and $|\ \mathbb{B}(\hat{V}(\hat{t}))\ |$. Tables 5.1 and 5.2 clearly show that the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}1np}$ are the best. These estimators are less sensitive to nonresponse if we consider the values of $\text{RE}(\hat{t})$ and $\text{RE}(\hat{V}(\hat{t}))$ (Table 5.3). Nevertheless the criterion of reliability of the confidence intervals ($\text{RB}(\hat{t}) < 0.10$) is never met (Table 5.4).

(iii) The behaviour of the estimators adjusted (i) for item y_1, which is the item the most highly correlated with the auxiliary variable, compared to item y_3, then (ii) for item y_2 compared to item y_3 (y_3 being the item that is least correlated with the auxiliary variable), shows that with very strong explanatory variables (for y_q and for Θ_{qk}), better results can be achieved not only in terms of less bias $|\ \mathbb{B}(\hat{t})\ |$ and $|\ \mathbb{B}(\hat{V}(\hat{t}))\ |$ but also in terms of less mean square error (a gain in precision in relation to the naive estimator) and a better recovery rate for the confidence intervals (Tables 5.1 to 5.4).

(iv) The behaviour of the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}np}$, in terms of bias, variance and variance estimation, is consistent with the studies conducted by Särndal and Hui (1981), Särndal and Swenson (1985, 1987), Bethlehem (1988) and Kott (1987) on the usefulness of regression estimators in nonresponse situations and the importance of having good predictor variables for the items of interest and the response mechanisms.

ACKNOWLEDGEMENTS

I wish to express my thanks to Carl-Erik Särndal for his support in every sense of the word in the writing of my Ph.D. thesis, on which this article is based. Despite his many responsibilities and the other demands on his time, he taught me a great deal in this field of sampling, which he masters so well and in which he has become a figure of international prominence through his many published works (articles and books) and collaborative efforts.

I would also like to thank the referees and the Associate Editor for their constructive comments. On the one hand, their observations and suggestions improved the original version of this article. On the other, they provided ideas for subsequent studies.

REFERENCES

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

CHICOINEAU, F., PAYEN, J.F., and THÉLOT, C. (1985). Modélisation et redressement des non-réponses: le cas du salaire. *Bulletin of the International Statistical Institute*, LI-3, 15.3, 1-23.

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

GIOMMI, A. (1985). On the estimation of the individual response probabilities. *Bulletin of the International Statistical Institute*, 2, 577-578.

GIOMMI, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology*, 13, 127-134.

GROSBRAS, J.-J. (1987b). Les réponses manquantes. In *Les sondages*. (Eds. J.-J. Droesbeke, B. Fichet and F. Tassi). Paris: Economica.

JOHNSON, N.L., and KOTZ, S. (1970). *Continous univariate distributions-1*. New York: Houghton.

KOTT, P.S. (1987). Nonresponse in a periodic sample survey. *Journal of Business and Economic Statistics*, 5, 287-293.

KRAFT, C.H., LEPAGE, Y., and VAN EEDEN, C. (1983). Some finite-sample size properties of Rosenblatt density estimates. *The Canadian Journal of Statistics*, 11, 95-104.

OH, H.L., and SCHEUREN, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184. New York: Academic Press.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of the density function. *Annals of Mathematical Statistics*, 27, 832-837.

SÄRNDAL, C.-E., and HUI, T-K. (1981). Estimation for nonresponse situations: to what extent must we rely on models? In *Current Topics in Survey Sampling*. (Eds. D. Krewski, R. Platek and J.N.K. Rao), 227-246. New York: Academic Press.

SÄRNDAL, C.-E., and SWENSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute*. LI-3, 15.2, 1-16.

SÄRNDAL, C.-E., and SWENSON, 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.

WEGMAN, E.J. (1972a). Nonparametric probability density estimation: A summary of available methods. *Technometrics*, 14, 533-546.

WEGMAN, E.J. (1972b). Nonparametric probability density estimation: A comparison of density estimation methods. *Journal of Statistical Computations and Simulations*, 1, 225-245.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 185–192
Statistics Canada

Competitors to Genuine πps Sample Designs: A Comparison

OLIVER SCHABENBERGER and TIMOTHY G. GREGOIRE[1]

ABSTRACT

Without-replacement list sampling with probability proportional to some measure of element size has not enjoyed much application in forestry because of the difficulty of implementing such sample strategies, that have been termed πps designs to distinguish without-replacement sampling from the well-known with-replacement pps designs. In this contribution, an exact πps strategy (Sunter's variant 2), an approximate πps design (Sunter's variant 1) and the Rao-Hartley-Cochran random group method are examined and the variances of the respective estimators for total bole volume are computed for four tree populations. The results indicate that compared to the Rao-Hartley-Cochran design Sunter's variant 1 in general leads to higher precision if the relationship between auxiliary information x_k and target characteristic y_k is *loose* but is sensitive to the ordering of the sampling frame, whereas the Rao-Hartley-Cochran design does not require the sampling frame to be ordered at all and appears to be superior if strong linear relationships between x_k and y_k are present.

KEY WORDS: Probability proportional to size sampling; Fixed sample size; Approximate πps designs; Empirical comparison.

1. INTRODUCTION

Rao (1978) classifies methods for unequal probability sampling without replacement in two broad categories, (i) sampling schemes, where the inclusion probabilities π_k are proportional to the characteristic of interest, y_k, and the Horvitz-Thompson π estimator $\hat{t}_\pi$ is utilized; (ii) schemes that entertain statistics other than the Horvitz-Thompson estimator. Strategies in (i) are termed IPPS (inclusion probability proportional to size) and members of (ii) non-IPPS designs. In recent literature, *e.g.*, Särndal *et al.* (1992), selection probabilities when sampling with-replacement are denoted p, whereas their counterparts when sampling without replacement are denoted π. We therefore call sampling designs in (i) genuine πps strategies in this paper. Both, IPPS and non-IPPS designs have in common, that under exact proportionality, *i.e.*, $\pi_k \propto y_k$ and $n(s) \equiv n\{\text{constant}\}$, it is implied that $\text{Var}(\hat{t}) \equiv 0$ where $\hat{t}$ is the respective estimator used. For this reason, it seems appealing to draw a sample without replacement where $\pi_k \propto y_k$ and to keep the sample size fixed at the same time. Our interest in these methods concerns their utility to sampling needs in forestry.

Several exact πps designs are available, Rao (1978) gives an in depth account and discussion. Their implementation however is often a non-trivial task and numerically cumbersome for sample sizes usually encountered in forestry practice. Many of these exact πps strategies require enumeration of all possible samples or use algorithms that become increasingly prohibitive as n increases. A simple design, which is feasible for $n \leq 10$ is described by Sampford (1967).

In forestry, however, the number of samples to be drawn at any stage of a survey is oftentimes much larger, even after stratification. Consequently, one either approximates the πps selection process in a manner that allows the inclusion probabilities to be computed exactly, or approximates second-order inclusion probabilities π_{kl} in a design that ensures an exact πps selection. Rao, Hartley and Cochran (1962) described a non-IPPS design, also known as the random group method, that has gained considerable attention (see also Rao 1966, 1978). It is not a πps design, since it utilizes an estimator other than $\hat{t}_\pi$ to ensure zero variance when the π_k are proportional to y_k, but is of remarkable simplicity. An approximate πps design of the first kind is Sunter's method (Sunter 1977a, 1977b). These two designs are referred to in what follows as RHC and SUN1. Sunter (1986, 1989) described an exact πps strategy that can be applied if certain stipulated conditions about the ordering of the sampling frame are met and the possible samples can be enumerated to obtain π_{kl} for some pairs of elements. To avoid enumeration we use an approximation to these π_{kl}. This scheme will be called variant 2 or SUN2 in what follows.

Särndal *et al.* (1992) describe the SUN1 and RHC strategies as entailing some loss of efficiency compared to corresponding πps designs, but no assessment of their comparative efficiency is provided. To our knowledge, none is extant; yet in light of the practical advantages offered by these designs, a comparative assessment would be helpful.

[1] Oliver Schabenberger and Timothy G. Gregoire, Department of Forestry, Section Forest Biometrics, College of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, U.S.A.

The purpose of this study is to compare the performance of the three strategies empirically, using data from forestry field studies and sampling intensities up to 10% which involve reasonably large samples.

The designs SUN1, SUN2, and RHC are appropriate if one has access to a list of population elements from which the sample can be drawn. A complete enumeration of the target characteristic y_k is not anticipated, but the probabilities of inclusion may be made proportional to an auxiliary variable x_k. That is, having complete knowledge about x_k prior to sampling, where it is surmised that x_k is roughly proportional to y_k, we try to achieve $\pi_k \propto x_k$ while $n \equiv$ constant.

In forestry such auxiliary information oftentimes is an easily obtainable characteristic of tree size such as height h, diameter at breast height d, or a combination thereof, which can be used to sample efficiently for bole volume or biomass, y. For example, the geometry of tree stems suggests relationships between d, h, and the volume contained in the tree bole that can be exploited in sampling. In the present investigation, the target parameter is the total bole volume per unit area or in an entire forest stand. In practice, some form of multistage sampling would be used, but for sake of exposition the present comparison includes single stage sampling only.

For the RHC and SUN designs, the auxiliary variables d, d^2, d^2h and the tree sequence number were used. The sequence number was chosen as an auxiliary variable since in the absence of ordering by size it is clearly unrelated to the target characteristic. It should indicate the sensitivity of competing strategies to uninformative auxiliary information (*cf.* Rao 1966).

All designs were investigated with samples of intensity 1%, 2%, 5%, and 10%. The performance of the different sampling designs was gauged in terms of the variance of each estimator of $t = \sum_{k=1}^{N} y_k$. Ratio-of-means estimation following simple random sampling was used as a benchmark, since it utilizes the same auxiliary information. The variances of the sample designs described in the following section were compared to the mean square error of the ratio-of-means estimator (ROM), evaluated using the second order delta method approximation in Sukhatme *et al.* (1984).

2. SAMPLE DESIGNS

2.1 Sunter's Design, Variant 1

Sunter initially proposed two different approximate πps designs: one relaxes the requirement of proportionality of inclusion probabilities π_k for a subset of the population, the other allows for some variation in sample size (Sunter 1977a, 1977b; Schreuder *et al.* 1990). In order that precision not be unduly sacrificed, it is assumed in the latter case that the variance of $n(s)$ is small, while in the first case that altering some π_k is not too serious. In this study only the first method was used since the RHC design operates with fixed sample size, too, and it is the comparative feasibility of the Sunter and RHC designs that prompted this study. Särndal *et al.* (1992) describe the allocation of the sample and the computation of the inclusion probabilities in detail. For part of the population, $\pi_k \propto x_k$ where x_k is the auxiliary information available for the k-th subject (or record). Let k^* denote an element in the ordered population. Then for all elements where $k < k^*$ selection is carried out proportional to x_k. The process ends if a total sample of size n is allocated or if $k = k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ where $t_k = \sum_{j \geq k} x_j$. In the latter case, the remaining samples are selected according to the list-sequential scheme of Bebbington (1975) among those elements for which $k \geq k^*$. As Sunter points out, this sampling scheme has the advantage that only one pass through the sampling frame is necessary. Moreover, the first and second order inclusion probabilities can be computed during this pass through the file. Since the design ensures that $\pi_{kl} > 0 \; \forall \; k, l$; $\pi_k\pi_l - \pi_{kl} > 0 \; \forall \; k, l$ and n is fixed, the non-negative Yates-Grundy estimator of variance can be readily computed. The first order inclusion probabilities are obtained as $\pi_k = nx_k/T_N$ if $k < k^*$ and $\pi_k = n\bar{x}_k/T_N$ if $k \geq k^*$ where $T_N = \sum_{k=1}^{N} x_k$ and $\bar{x}_{k^*} = t_{k^*}/(N - k^* + 1)$. Expressions for the second order inclusion probabilities are given in Särndal *et al.* (1992).

Consequently, the ordering of the population affects the performance of the SUN1 design, since the inclusion probabilities and therefore the variance depend on k^* (see (2) below). For large sample sizes the condition $k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ may be resolved in favor of $k^* = \min\{k: nx_k/t_k \geq 1\}$, which in turn may lead to a premature switch from πps to SRS sampling owing to the ordering of the sampling frame. Note that $x_k/t_k < x_{k'}/t_{k'}$ for $k' > k$ need not be true since if $x_k > x_{k+1}$ and $t_k > t_{k+1}$ it may well be that x_k/t_k is greater or smaller than x_{k+1}/t_{k+1}. It thus can happen that $nx_k > t_k$ and $nx_{k'} < t_{k'}$, for some k, k' where $k' > k$. In this case, that may occur rather frequently, it is unclear if the switch from πps to SRS should take place the first time $nx_k \geq t_k$ or not. Sometimes it may happen that for the first two or three elements of the population $nx_k \geq t_k$ but falls below t_k for the main portion of the sampling frame. This is especially the case when n is large and a few very big x_k appear on top of the population list. To stick to Sunter's rule in such a case would in essence be equivalent to drawing a simple random sample.

The π estimator for the population total can be computed as

$$\hat{t}_{\pi\text{SUN1}} = \sum_{k=1}^{N} \frac{y_k}{\pi_k} I_k, \qquad (1)$$

where I_k is the sample inclusion indicator function. The variance is obtained as

$$\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN1}}) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\mathrm{Cov}(I_k,I_l)\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2, \quad (2)$$

which is the Yates-Grundy form with $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ (Särndal *et al.* 1992). We use the notation $\mathrm{VAR}_{\mathrm{SUN1}}$ for (2) subsequently.

2.2 Sunter's Variant 2

In Sunter (1986, 1989) an exact πps design is described for samples of size $n > 2$. To fix ideas let $z_k = x_k/T_N$ and order the population such that

$$nz_k < Z_k,\ k = 1, \ldots, N - (n+1)$$

$$(n-k)z_l < Z_k,\ l \geq k \geq N - n,$$

where $Z_k = \sum_{i=k}^{N} z_i$. Let m_k denote the number of samples out of n still to be drawn when arriving at the k-th population element u_k. Given that the two conditions are met, the following algorithm selects an exact πps sample. For u_k, $P(u_k \mid m_k) = nz_k/Z_k$ until $m_k = 0$ or $m_k = N - k$; in the latter case discard one of the remaining units with probability $1 - (m_k z_l/Z_k)$ and retain the others.

It is not always possible to order the population such that the above conditions are met. Sunter (1986) describes an algorithm that checks, whether the ordering is possible. The inclusion probabilities are

$$\begin{aligned}\pi_k &= nz_k \\ \pi_{kl} &= n(n-1)z_k z_l \gamma_k \quad k \leq N - n - 1,\ l > k,\end{aligned} \quad (3)$$

where

$$\gamma_k = \frac{1}{Z_{k+1}}\left(1-\frac{z_1}{Z_2}\right)\cdots\left(1-\frac{z_{k-1}}{Z_k}\right), \quad k = 2, \ldots, N-(n+1).$$

The remaining second-order inclusion probabilities, namely π_{kl} for $l > k > N - n$ have to be obtained from enumeration of possible samples which is likely to be infeasible. Sunter argues that (3) gives a good approximation for those pairs of elements, and this approximation has been used here. With these inclusion probabilities, $\hat{t}_{\pi\mathrm{SUN2}}$ is indicated by the right-hand-side (rhs) of (1). An approximation to $\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN2}})$ is given by (2), wherein (3) is used to obtain π_{kl} for $l > k > N - n$.

The differences between SUN1 and SUN2 are noteworthy. With SUN1 the joint inclusion probabilities are computed exactly for all pairs, but the selection is not genuine πps because of the introduction of SRS in part. In Sunter's variant 2 the selection is exactly πps, but $\mathrm{Var}(\hat{t}_{\pi\mathrm{SUN2}})$ can only be approximated. We use $\mathrm{VAR}_{\mathrm{SUN2}}$ to denote this approximation.

2.3 RHC Design

A description of the RHC design is straightforward; properties of the RHC estimator are well documented in Rao, Hartley and Cochran (1962), and Rao (1966, 1978). After fixing the sample size n, the universe of size N is randomly divided into n groups of size N_i where $N = \sum_i N_i$ $(i = 1, \ldots, n)$. Let X_{ik} denote auxiliary information for element u_k in group i, $k = 1, \ldots, N_i$, and put $X_{i.} = \sum_{k=1}^{N_i} X_{ik}$. From each group one element is selected with selection probability $p_{ik} = X_{ik}/X_{i.}$. The estimator for the total in group i is given as

$$\hat{t}_{i\pi} = \sum_{k=1}^{N_i}\frac{y_{ik}}{p_{ik}}I_{ik},$$

where I_{ik} is the sample inclusion indicator function for element u_k in group i. The population total is then estimated by

$$\hat{t}_{gr} = \sum_{i=1}^{n}\hat{t}_{i\pi}, \quad (4)$$

with variance

$$\mathrm{Var}(\hat{t}_{gr}) = \frac{1}{N(N-1)}\left(\sum_{i=1}^{n}N_i^2 - N\right)\left(\sum_{k=1}^{N}T_N y_k^2/x_k - t^2\right). \quad (5)$$

Note that (5) depends on the group sizes and is minimized when all are equal. In our application, we determined N_i such that some groups were of size $N_i = [N/n]_{gif}$ where *gif* denotes the greatest integer function and the remainder of size $N_i = [N/n]_{gif} + 1$. The number of groups of each size is chosen so that the sum of the group sizes is N. If N/n is an integer, all groups are of course of equal size. We denote (5) by $\mathrm{VAR}_{\mathrm{RHC}}$ in the sequel.

The RHC design is not an exact πps design, since the subdivision of the population introduces a source of randomness unrelated to the size of the auxiliary variable and (4) is not a Horvitz-Thompson estimator. The inclusion probability depends jointly on the size of X_{ik} and on the probability of an element being assigned to group i. Ordering of the population has no effect on $\mathrm{VAR}_{\mathrm{RHC}}$.

3. TREE POPULATIONS

Table 1 shows the tree populations under consideration and Figure 1 displays the relationship between the various choices for x_k and the target characteristic for the yellow poplar population. We notice almost perfect proportionality between d^2h and volume, the relationship between d and volume is clearly curvilinear, and the relationship between d^2 and volume is intermediate. No noticeable trend between sequence number and volume is apparent in the unordered sampling frame. For the remaining three populations similar patterns hold.

For the four populations and the various combinations of auxiliary variable and sampling intensity, there were no observations for which $nx_k > T_N$, thus no records were measured with certainty.

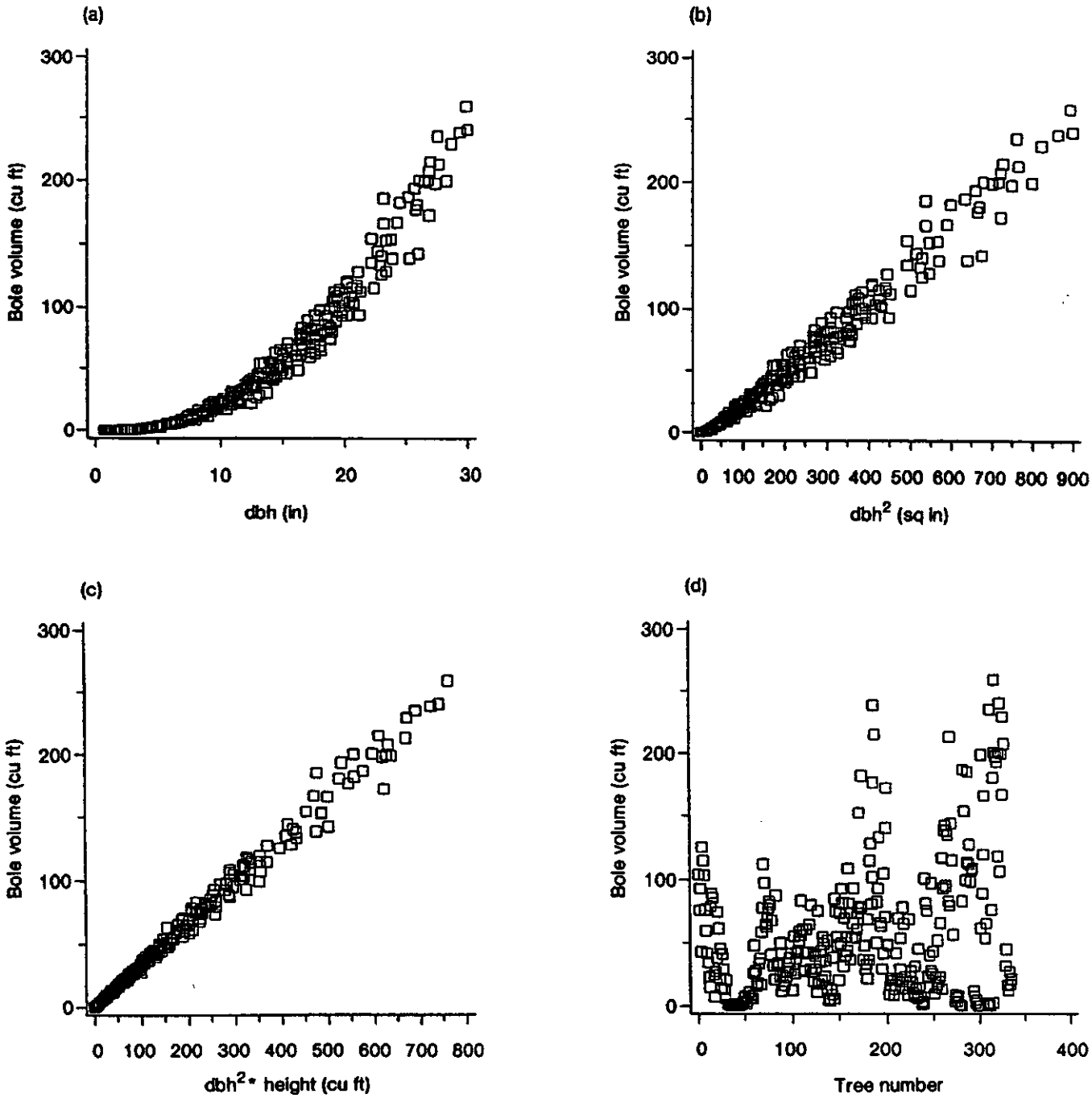

Figure 1. Relation of bole volume to bole dimensions in yellow poplar: (a) diameter at breast height; (b) diameter squared; (c) squared diameter times height; (d) tree sequence number.

Table 1

Tree Populations Examined in an Empirical Comparison of SUN1, SUN2, and RHC
The Last Four Columns Contain Pearson Correlation Coefficients Between x_k and y_k

Species		$N^{(1)}$	$t(ft^3)^{(2)}$	$\rho_{(y;x)}$			
				d	d^2	d^2h	No
Ponderosa pine	*Pinus ponderosa*	140	9,366.6	0.99	0.99	0.99	0.31
Yellow poplar	*Liriodendron tulipifera*	336	18,255.5	0.96	0.96	0.99	−0.07
Loblolly pine	*Pinus taeda*	437	1,835.8	0.96	0.96	0.99	−0.32
Red pine	*Pinus resinosa*	91	4,075.7	0.96	0.96	0.97	−0.05

(1) N is the number of trees in the population.
(2) t is total volume.

4. RESULTS

4.1 Comparison of Variances

The variance of the estimators of t corresponding to the SUN1, SUN2, and RHC design, expressed as a proportion of the MSE under the ROM strategy are compared in Table 2 for the yellow poplar population for each of the sampling intensities investigated and Table 3 depicts pertinent results for the remaining populations. For the SUN1 strategy, the populations were ordered by decreasing size of X, as recommended by Sunter (1977a, 1977b). We focus initially on the results for the yellow poplar population in Table 2.

Table 2

Relative Performances of SUN1, SUN2 and RHC Design for the Yellow Poplar Population where Ratio-of-means Estimation (ROM) Serves as a Benchmark

$n/N\%$	X	n	$\frac{VAR_{SUN2}}{MSE_{ROM}}$	$\frac{VAR_{SUN1}}{MSE_{ROM}}$	$\frac{VAR_{RHC}}{MSE_{ROM}}$	k^{*1}
1	*No*	4	4.8120	3.3136	4.7767	
1	d	4	0.6735	0.6684	0.6731	332
1	d^2	4	0.4605	0.4596	0.4613	333
1	d^2h	4	0.3361	0.3378	0.3402	330
2	*No*	7	5.1327	2.6346	5.0568	
2	d	7	0.7090	0.6982	0.7081	325
2	d^2	7	0.5731	0.5694	0.5751	318
2	d^2h	7	0.4263	0.4542	0.4369	316
5	*No*	17	5.4938	1.6643	5.2793	
5	d	17	0.7305	0.7808	0.7283	309
5	d^2	17	0.6541	0.6992	0.6608	291
5	d^2h	17	0.4603	1.2638	0.4935	285
10	*No*	34	5.8326	1.0985	5.3594	
10	d	34	0.7385	0.7083	0.7339	247
10	d^2	34	0.6712	0.9687	0.6864	260
10	d^2h	34	0.4298	3.0140	0.5037	250

[1] k^* is the observation in the ordered sampling frame at which the SUN1 design switches from πps to SRS sampling.

For a given sampling intensity the precision of all designs relative to ROM increases in the order $X \equiv No$, d,d^2,d^2h; *i.e.*, with increasing proportionality between auxiliary variable and tree bole volume. Given that the approximation of the variance of SUN2 performs well, VAR_{SUN2} can be regarded as measuring the closeness of the RHC and SUN1 designs to matching the efficiency of a genuine πps selection. At low sampling intensities and with meaningful auxiliary information the two designs do not deviate much from SUN2. The performance of both RHC and SUN1 appears to deteriorate at higher sampling intensities relative to SUN2 depending on the choice of size measure. For $X \equiv d^2h$, in which case $\rho_{(y;x)} \cong 0.99$ (see Table 1), RHC is still .85 (.4298/.5037) as efficient as SUN2 but SUN1 is only .14 (.4298/3.014) as efficient, when $n/N\% = 10$. The performance of RHC and SUN1 relative to SUN2 improves for other choices of X which are less well correlated with Y. Indeed, when $X = No$, SUN1 is much more efficient than SUN2.

A puzzling aspect of these results is the indication that SUN2 is less efficient than either RHC or SUN1 for some choices of auxiliary variable and sampling intensity. We speculate that it may be an artifact of the approximation of some second-order inclusion probabilities incorporated into VAR_{SUN2}. It also may depend on the particular ordering used in SUN1 or the group sizes used in RHC sampling, respectively. It is feasible to calculate the exact $Var(\hat{t}_{\pi SUN2})$ for $n = 2$. We did so for the ponderosa pine and the red pine populations. The results indicate that VAR_{SUN2} approximates the precision of the SUN2 design very well, but is slightly conservative. The ratios $Var(\hat{t}_{\pi SUN2})/VAR_{SUN2}$ took on values between 0.975 and 0.999. For larger sample sizes there is no feasible way to determine how well the approximation VAR_{SUN2} performs.

We focus now on the comparison of RHC to SUN1, again with reference to Table 2. At low sampling intensities, VAR_{SUN1} and VAR_{RHC} are essentially equivalent when $X \equiv d^2h$. But using this auxiliary variable at higher intensities led to a substantially better performance of $\hat{t}_{gr}$ in some cases. The most noteworthy case is $n/N\% = 10$ where $\hat{t}_{gr}$ is nearly 6 times more precise than $\hat{t}_{\pi SUN1}$.

We surmise from these results that the better $x_k \propto y_k$ holds, the better is the precision of $\hat{t}_{gr}$ relative to $\hat{t}_{\pi SUN}$ owing chiefly to the effect of k^* on VAR_{SUN}. Small values of k^* indicate an early switch to a SRS selection and coincide with small values of VAR_{SUN2}/VAR_{SUN1}. Large values of k^* on the other hand correspond to variance ratios close to 1. For yellow poplar, $n/N\% = 10$ and $X \equiv d^2h$ the SUN1 design selects only three-fourths of the population according to a πps design; we conjecture that the early transition to SRS serves also as an explanation for its poor performance compared to the RHC design. When X = tree sequence number, SUN1 is much more precise than RHC, and its relative precision increases as n increases.

The sharp improvement in efficiency when using an auxiliary variable other than tree sequence number provides an indication of the effectiveness of the strategies discussed here when X is positively correlated to Y, and to the liability of sampling with probability proportional to an auxiliary variable when it is unrelated to Y.

The pattern evident in the results for yellow poplar are generally seen, also, in the results for the other species. Some of them are summarized in Table 3. For ponderosa pine SUN1 relative to RHC is always less precise when $X \equiv d^2h$ regardless of the sampling intensity and SUN2 performs always best when this variable is used. For all species the combination $n/N\% = 10$, $X \equiv d^2h$ leads to low precision of SUN1 compared to the other designs and with the exception of the loblolly pine population, SUN1 performs poorer than ratio-of-means estimation. For all populations, the order of magnitude better precision of ROM over the genuine πps, non-IPPS or approximate πps design when X = tree sequence number is remarkable.

From Figure 1 it can be seen that the ordering of volume by tree numbers is haphazard, *i.e.*, the sequence number carries no information about bole volume. And, there is a price to pay if one uses this uninformative auxiliary information to determine inclusion probabilities. The inefficiency of unequal probability sampling in presence of uninformative auxiliary information is an important limitation for the simultaneous estimation of multiple population attributes, where some may be closely related to the auxiliary design variable but others might be uncorrelated with it. Rao (1966) discusses this point in detail and he proposes alternative estimators based on the unbiased estimators in equal probability sampling and the estimator $\hat{t}_{gr(alt)} = N\sum_i y_i \xi_i$, where $\xi_i = \sum_k^{N_i} p_{ik}$ in the RHC design. Applying this estimator in the case of unequal probability sampling leads to bias, but to better mean-square error performance. For the RHC design with X = tree sequence number, the alternative estimator proposed by Rao (1966) improved the ratio $MSE_{RHC(alt)}/MSE_{ROM}$ remarkably. For the yellow poplar population for example, these ratios were between 1.34 ($n = 4$) and 2.58 ($n = 34$), corresponding to a mean square error of the alternative estimator of only 28% to 48% ($n = 34$) of the RHC estimator (5). Similar patterns hold for the other tree species.

Since the alternative estimator is inconsistent, its bias does not depend on n, the larger ratios within the range for each species appear for larger sample sizes. It thus seems reasonable to limit the use of this estimator to smaller sample sizes. When n gets larger, another alternative is to use a ratio estimator, *e.g.*, Hajek's estimator $N\{(\sum y_i/\pi_i)/(\sum 1/\pi_i)\}$ under a genuine πps design.

Table 3

Pertinent Results About the Relative Performances of SUN1, SUN2 and RHC Design for the Remaining Populations where Ratio-of-means Estimation (ROM) Serves as a Benchmark

$n/N\%$	X	n	$\frac{VAR_{SUN2}}{MSE_{ROM}}$	$\frac{VAR_{SUN1}}{MSE_{ROM}}$	$\frac{VAR_{RHC}}{MSE_{ROM}}$	k^*
			Ponderosa Pine			
1	*No*	2	1.9608	1.9794	1.9507	
1	d^2h	2	0.1050	0.1096	0.1077	137
2	*No*	3	2.2976	1.9264	2.2275	
2	d^2h	3	0.1768	0.1919	0.1859	135
5	*No*	7	2.8717	2.0681	2.7819	
5	d^2h	7	0.3113	0.3890	0.3670	129
10	*No*	14	3.2528	2.2745	3.0294	
10	d^2h	14	0.2928	1.3724	0.4488	97
			Red Pine[1]			
2	*No*	2	2.0210	1.9485	2.0029	
2	d^2h	2	0.9076	0.9026	0.9104	90
5	*No*	5	2.9295	2.3141	2.8236	
5	d^2h	5	0.8874	1.3456	0.8991	87
10	*No*	9	3.5548	2.0124	3.2958	
10	d^2h	9	0.8699	1.3192	0.8942	81
			Loblolly Pine			
1	*No*	5	4.8011	3.7104	4.7625	
1	d^2h	5	0.4043	0.4161	0.4174	431
2	*No*	9	5.5940	3.7441	5.5044	
2	d^2h	9	0.5129	0.5510	0.5476	419
5	*No*	22	6.5290	3.3082	6.5253	
5	d^2h	22	0.5035	0.6385	0.6085	406
10	*No*	44	7.7977	2.6635	6.5708	
10	d^2h	44	0.3854	0.7214	0.6146	375

[1] The sampling intensity 1% was omitted since it would have resulted in $n = 1$.

4.2 The Effect of Ordering on The Precision of Sunter's Variant 1

Sunter and others have noted that the precision of the SUN1 design depends on the ordering of the population. The recommendation to sort the sampling frame by decreasing size of x_k's is rooted in the assumption that larger x_k are more likely to be proportional to y_k than smaller ones. The goal is to apply the πps part of the SUN1 design not only to as big a portion of the population as possible but also to those elements for which $x_k \propto y_k$ holds best. Under this assumption it was thus advised to put the elements with large x_k values at the top of the frame. However, it is clear that this is only a rough rule of thumb, since the assumption of greater proportionality with increasing size may not hold.

To investigate the effect of ordering the ponderosa pine and red pine populations were first sorted by increasing x_k and then grouped into 10 groups of approximately equal size. The Pearson correlation coefficient between x_k and y_k was computed within each group and the populations were then sorted by

(a) groups of decreasing correlation and increasing size of x_k within each group,

(b) groups of decreasing correlation and decreasing size of x_k in each group

and SUN1 sampling was repeated for the combinations of x_k's and sampling intensity 10%. Table 4 shows the results.

Table 4

$\text{Var}_{\text{SUN1}}/\text{MSE}_{\text{ROM}}$ for Ponderosa Pine and Red Pine and Different Ways of Ordering the Population

	Ponderosa Pine Ordered by			Red Pine Ordered by		
X	decr.x_k	decr.ρ incr.x_k	decr.ρ decr.x_k	decr.x_k	decr.ρ incr.x_k	decr.ρ decr.x_k
d	0.5614	0.6165	0.6043	1.0307	1.0236	0.6454
d^2	0.3478	0.6562	0.5869	1.2077	0.9373	0.6948
d^2h	1.3724	60.861	0.4459	1.3192	0.8674	0.7461

The results are rather surprising. For red pine the order by decreasing correlation improved all measures of precision. Sorting by increasing x_k within each group now made VAR_{SUN1} very close to VAR_{RHC}, and with $x = d^2h$, $\text{VAR}_{\text{SUN1}} < \text{VAR}_{\text{RHC}}$. Sorting by decreasing x_k within each group achieved an even greater improvement. In contrast to these results, sorting the ponderosa pine population by decreasing ρ and increasing x_k made things worse. The very high value of 60.861 is caused by a premature switch to SRS, since in this setting k^* is only 28, corresponding to only 20% of the population being sampled πps. Moreover, using order of decreasing ρ and decreasing x_k improved VAR_{SUN1} only for $x = d^2h$.

These results indicate that there may exist an order that minimizes VAR_{SUN1} and may yield higher precision than a simple ordering by decreasing value of X. But this order will usually differ depending upon the auxiliary information, and even an ordering that is reasonable on intuitive grounds may give unanticipated results. It is not known if any ordering is optimal in the sense of minimizing $\text{Var}(\hat{t}_{\pi\text{SUN1}})$ for the approximate πps design used in this study. According to our present knowledge no optimal strategy has been described.

5. DISCUSSION AND CONCLUSION

Employing some meaningful auxiliary information leads to a considerable gain in precision in the unequal probability designs compared to a ratio-of-means estimation.

A choice between the two Sunter designs can be made on grounds of the relationship between size measure and target characteristic. When $X \propto Y$ is strong, SUN2 offers advantage over SUN1, and SUN1 appears preferable when the relationship is weak. Based on our results, the approximate πps strategy, SUN1 and the non-IPPS design RHC appear to come fairly close to the efficiency offered by genuine πps selection. With increasing sampling intensity, however, the highest precision is obtained with the SUN2 design. But the quality of the approximation VAR_{SUN2} in this case is unclear.

If one's aim is to use an approximate πps or a non-IPPS strategy then the RHC design with estimator $\hat{t}_{gr}$ appears to offer advantages over the Sunter design with $\hat{t}_{\pi\text{SUN}}$, at least for the tree populations studied here with the objective of estimating total bole volume. At reasonably low sampling intensities, both estimators appear to be equally precise.

An advantage of the RHC design is its simplicity. An operational advantage is that it can be applied to every population because it is impervious to its ordering and provides an unbiased estimation within each group. While the first criterion is also met by Sunter's variant 1, the ordering there clearly affects the precision of the estimator $\hat{t}_{\pi\text{SUN1}}$. Variant 2 can only be used if some ordering of the population meets the conditions given in Section 2.2. Otherwise the selection algorithm does not produce a sample of exactly size n.

The precision of the RHC method, however, depends on the group sizes employed. The algorithm given in Section 2.3 is optimal.

While a particular ordering may improve the precision of $\hat{t}_{\pi\text{SUN1}}$, it is unclear at present how to discern an optimal ordering and a fixed sample size. Moreover an optimal

ordering of one choice of auxiliary variable or attribute of interest may be deleterious when implemented with a different auxiliary variable or attribute.

All strategies can be disastrous with uninformative auxiliary information.

Finally and to the extent that computational burden is a meaningful criterion, RHC is arguably less burdensome than variant 1 of Sunter's design.

ACKNOWLEDGMENT

We gratefully acknowledge the comments and suggestions by J.N.K. Rao, C.-E. Särndal, and A. Sunter who reviewed earlier versions of the manuscript as well as the helpful comments of the referees whose contribution helped to improve the paper substantially.

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā* A, 28, 47-60.

RAO, J.N.K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. *Contributions to Survey Sampling and Applied Statistics* (H.A. David, Ed.), New York: Academic Press, 69-86.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society* B, 24, 482-491.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

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

SUKHATME, P.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications (3rd Ed.)*. Iowa State University Press.

SUNTER, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.

SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.

SUNTER, A. (1989). Updating size measures in a PPSWOR design. *Survey Methodology*, 15, 253-260.

SCHREUDER, H.T., LI, H.G., and SADOOGHI-ALVANDI, S.M. (1990). Sunter's pps Without Replacement Sampling as an Alternative to Poisson Sampling. USDA Forest Service Research Paper RM-290.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1994. An asterisk indicates that the person served more than once.

M. Bankier, *Statistics Canada*
* G.E. Battese, *University of New England, Australia*
Y. Beaucage, *Statistics Canada*
* D.R. Bellhouse, *University of Western Ontario*
N. Bennett, *Yale University*
J. Bethel, *Westat, Inc.*
* J. Bethlehem, *Central Bureau of Statistics, The Netherlands*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistics Canada*
R.L. Chambers, *Australian National University*
S. Cheung, *Statistics Canada*
* N. Chinnappa, *Statistics Canada*
G.H. Choudhry, *Statistics Canada*
M.P. Cohen, *U.S. National Center for Education Statistics*
* M.J. Colledge, *Statistics Canada*
L.H. Cox, *U.S. Environmental Protection Agency*
C.Z.F. Clark, *U.S. Department of Agriculture*
R. Cochran, *University of Wyoming*
F. Conrad, *U.S. Bureau of Labor Statistics*
N. Cressie, *Iowa State University*
* J.-C. Deville, *Institut National de la Statistique et des Études Économiques*
P. Dick, *Statistics Canada*
D. Dillman, *Washington State University*
D. Dolson, *Statistics Canada*
* J.D. Drew, *Statistics Canada*
* F. Dupont, *Institut National de la Statistique et des Études Économiques*
W.S. Edwards, *Westat, Inc.*
E.P. Ericksen, *Temple University*
R.E. Fay, *U.S. Bureau of the Census*
* W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M. Ghosh, *The University of Florida*
* M. Gonzalez, *U.S. Office of Management and Budget*
H. Gough, *Statistics Canada*
* R.M. Groves, *University of Maryland*
* J.-P. Gwet, *Statistics Canada*
K.P. Hapuarachchi, *Statistics Canada*
H. Hogan, *U.S. Bureau of the Census*
* D. Holt, *University of Southampton*
A.Z. Israëls, *Central Bureau of Statistics, The Netherlands*
R. Jamieson, *Statistics Canada*
W.D. Kalsbeek, *University of North Carolina - Chapel Hill*
* G. Kalton, *Westat, Inc.*
* P.S. Kott, *National Agricultural Statistics Service*
* J. Kovar, *Statistics Canada*
* P. Lahiri, *University of Nebraska - Lincoln*
P. Lavallée, *Statistics Canada*
H. Lee, *Statistics Canada*
J.M. Lepkowski, *University of Michigan*
J.T. Lessler, *Batelle*
N.Y. Luther, *East-West Center*
P. Lys, *Statistics Canada*
D. Malec, *National Centre for Health Statistics*
* H. Mantel, *Statistics Canada*
M. March, *Statistics Canada*
H. Mariotte, *Institut National de la Statistique et des Études Économiques*
* A. Mason, *East-West Center*
N. Mathiowetz, *Agency for Health Care and Research*
P. Miller, *Narthurstan*
B. Nandram, *Worcester Polytechnic Institute of Mathematical Sciences*
J. Nealon, *National Agricultural Statistics Service*
J. Neter, *University of Georgia*
* D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
M. Ramos, *U.S. Bureau of the Census*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
L. Rizzo, *Westat, Inc.*
G. Roberts, *Statistics Canada*
K. Rust, *Westat, Inc.*
* I. Sande, *Bell Communications Research, U.S.A.*
* C.-E. Särndal, *Université de Montréal*
J. Schafer, *Pennsylvania State University*
* W.L. Schaible, *U.S. Bureau of Labor Statistics*
* F.J. Scheuren, *George Washington University*
I. Schiopu-Kratina, *Statistics Canada*
* J. Sedransk, *State University of New York*
A.C. Singh, *Statistics Canada*
* C.J. Skinner, *University of Southampton*
E.A. Stasny, *The Ohio State University*
T.W.F. Stroud, *Queen's University*
C.M. Suchindran, *University of North Carolina*
J. Tanur, *State University of New York - Stony Brook*
R. Tourangeau, *National Opinion Research Center*
R. Treder, *Statistical Sciences Inc.*
M.E. Thompson, *University of Waterloo*
J. Tourigny, *Statistics Canada*
* R. Valliant, *U.S. Bureau of Labor Statistics*
K.W. Wachter, *University of California - Berkeley*
* J. Waksberg, *Westat, Inc.*
M. Weekr, *Research Triangle Institute*
G.C. White, *Colorado State University*
W.E. Winkler, *U.S. Bureau of the Census*
* K.M. Wolter, *National Opinion Research Center*
* T. Wright, *Oak Ridge National Laboratory*
E. Zanutto, *Harvard University*
* A. Zaslavsky, *Harvard University*
J.V. Zidek, *University of British Columbia*

Acknowledgements are also due to those who assisted during the production of the 1994 issues: S. Beauchamp (Photocomposition) and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M.M. Kent, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 43, No. 4, 1994

Printed in Great Britain at the Alden Press, Oxford
This journal is printed on acid-free paper

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.